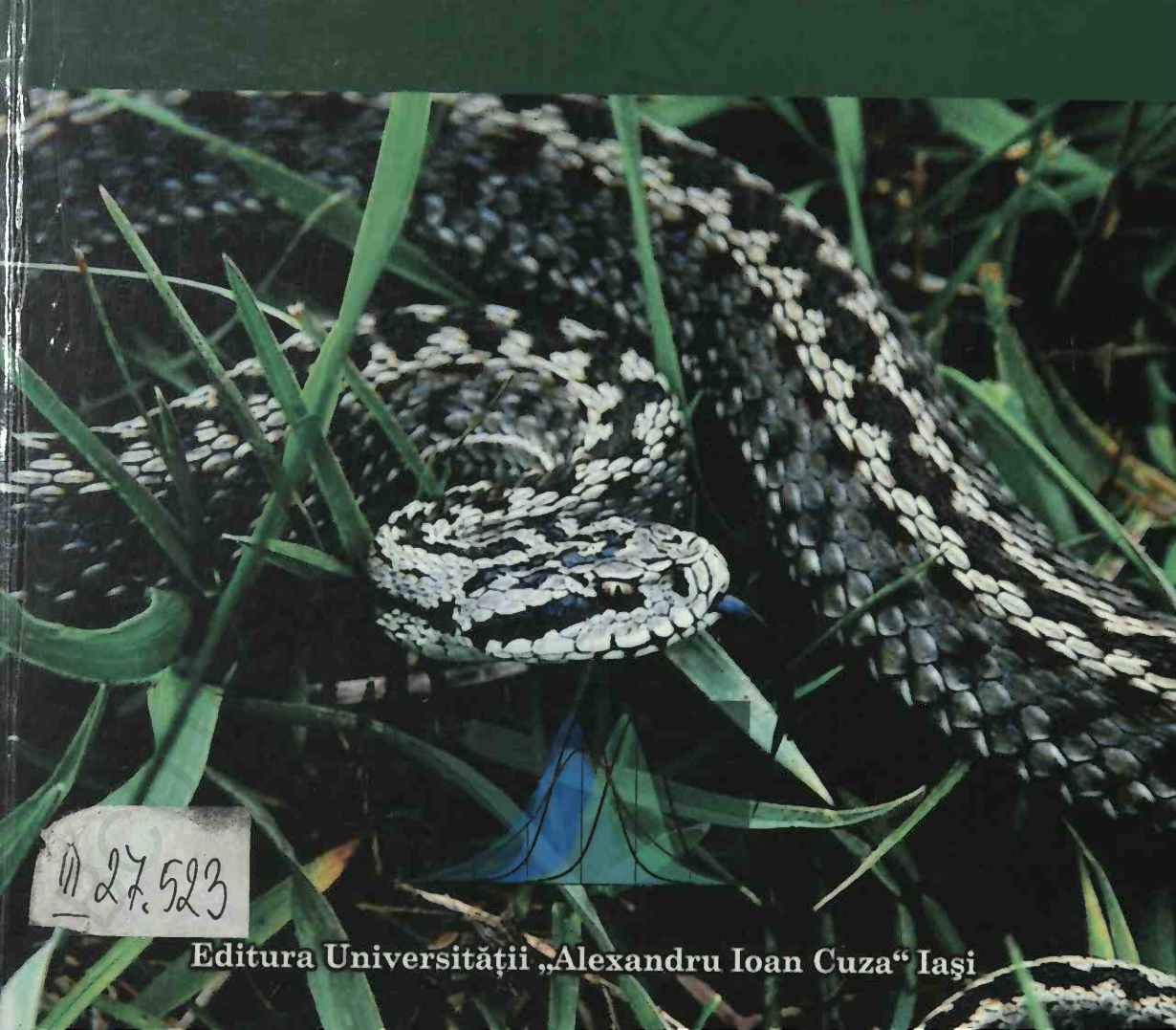
Ştefan R. Zamfirescu
Oana Zamfirescu

ELEMENTE DE STATISTICĂ APLICATE ÎN ECOLOGIE



Editura Universităţii „Alexandru Ioan Cuza" Iaşi

Ştefan Zamfirescu Oana Zamfirescu

Elemente de statistică aplicate în ecologie

Carte finanțată din Contract CNCSIS 1179/2006

Referenți științifici:

Prof.dr. Ioan Moglan
Departamentul de Biologie, Universitatea „Al.I.Cuza" Iași

Conf.dr. Luminița Bejenaru
Departamentul de Biologie, Universitatea „Al.I.Cuza" Iași

Lect.dr. Marcel Roman
Catedra de Matematică, Universitatea Tehnică „Gh. Asachi" Iași

Coperta: Manuela Oboroceanu
Redactor: Emil Juverdeanu

ISBN 978 - 973 - 703 - 389 - 5

700511 Iași, str. Păcurari nr. 9, tel./fax: 0232–314947

Ştefan Zamfirescu Oana Zamfirescu

Elemente de statistică aplicate în ecologie

EDITURA UNIVERSITĂŢII „ALEXANDRU IOAN CUZA“
IAŞI – 2008

Descrierea CIP a Bibliotecii Naționale a României
ZAMFIRESCU, ȘTEFAN
Elemente de statistică aplicate în ecologie/Ștefan R. Zamfirescu, Oana Zamfirescu - Iași : Editura Universității „Al. I. Cuza", 2008

ISBN 978 – 973 – 703 – 389 – 5

I. Zamfirescu, Oana

311:504(498)

CUPRINS

INTRODUCERE

În prezent, valorificarea investigațiilor ecologice nu poate fi concepută fără o analiză statistică a datelor, fără așa-numita asigurare statistică ce redă măsura în care concluziile acestor investigații ar putea fi reale. Analiza statistică a datelor, fără a fi un scop în sine al demersului științific în ecologie, reprezintă o unealtă ce permite o mai bună comprehensiune și prezentare a informației conținute de rezultatele cercetărilor.

În prezent, prelucrarea statistică a datelor este facilitată de utilizarea computerului și a programelor corespunzătoare. Utilizarea acestora trebuie să fie făcută numai după înțelegerea conceptelor și procedurilor metodelor statistice. Altfel, aceste instrumente vor reprezenta un fel de „cutie neagră" în care se introduc rezultatele cercetărilor și din care rezultă niște concluzii despre a căror corectitudine nu se poate spune mare lucru.

Prezenta lucrare încearcă să ofere o bază conceptuală pentru cei care se dedică cercetărilor cu caracter ecologic. Din acest motiv, pe parcursul capitolelor apar numeroase exemple inspirate din cercetări ecologice. Lucrarea poate fi însă de ajutor și pentru analiza datelor rezultate în urma investigațiilor din diverse ramuri ale biologiei sau ale științelor conexe.

În prima parte a cărții sunt prezentate o serie de noțiuni de bază în statistică, care să asigure înțelegerea limbajului intrinsec al acestei științe. Urmează o parte dedicată statisticii descriptive și principalelor distribuții probabilistice cu aplicativitate în ecologie. Partea următoare tratează aspectele de principiu ale statisticii inductive și prezintă principalele teste care se utilizează pentru comparația probelor. În continuare sunt prezentate modalități de analiză a asocierii dintre două variabile, iar ultima parte cuprinde metode de analiză ale frecvențelor. La finalul cărții sunt trei anexe: prima reprezintă o cheie de decizie asupra metodelor statistice, prin care pot fi prelucrate date, a doua cuprinde tabelele cu valorile critice necesare pentru diferite teste statistice, iar a treia cuprinde funcțiile statistice din programul MS-Excel. Cartea se termină cu un index alfabetic al termenilor statistici folosiți pe parcursul lucrării.

1. CONCEPTE GENERALE

Statistica reprezintă o parte importantă din preocupările actuale ale biologilor şi ecologilor. Termenul „statistică" este folosit în două ipostaze: fie se referă la colecţii de informaţie cantitativă şi la metode de procesare a acestora, fie la procesul de stabilire a unor concluzii privind grupuri de dimensiuni mari, în urma analizei unor părţi din aceste grupuri.

Statistica este ştiinţa care se ocupă cu organizarea, descrierea şi analiza numerică a fenomenelor cuantificabile, dezvăluind particularităţile lor de volum, structură, dinamică, conexiune, precum şi regulile sau legile care le guvernează.

Pentru ecologi şi pentru cei care, în general, studiază fenomene variabile cu implicaţii preponderent probabilistice, statistica este utilă pentru dirijarea colectării, organizării şi prezentării datelor, precum şi pentru tragerea concluziilor cu o anumită probabilitate sau grad de incertitudine de pe urma analizei datelor.

Trebuie menţionat că o analiză statistică nu demonstrează nimic, ci doar indică probabilitatea unui anumit rezultat sau concluzii trase în urma analizei datelor.

Atât în statistică, cât şi în cadrul ecologiei şi a celorlalte ramuri ale biologiei, apare noţiunea de **populaţie**. Accepţiunea biologică a acestui termen este de grup de indivizi ce aparţin unei anumite specii, între care se stabilesc interacţiuni şi ale căror gene alcătuiesc un genofond omogen. Din punct de vedere statistic, populaţia are un înţeles mai larg decât cel biologic, referindu-se la o colecţie de unităţi individuale, ce constituie obiectul unei investigaţii. Populaţia statistică reprezintă un grup de entităţi de un anumit tip din univers sau dintr-o subdiviziune specificată a universului. Este grupul de dimensiuni mari pe care dorim să-l cunoaştem. Aşa cum spuneam în primul paragraf, cunoaşterea unui astfel de grup sau populaţii se poate face prin intermediul analizei unei părţi. O astfel de parte, care este extrasă din populaţie pentru a fi studiată, se numeşte în general eşantion sau **probă**. Noi vom folosi în continuare noţiunea de probă bine încetăţenită în

comunitatea științifică ecologică. Deci proba este un grup mai mic, dar reprezentativ pentru populația din care a fost extras.

Studiul unei populații presupune investigarea uneia sau mai multor caracteristici ale unităților din probe, caracteristici care se numesc **variabile**. Valorile unei variabile corespunzătoare entităților unei populații se numesc **valori individuale**. Valorile individuale cunoscute, corespunzătoare unităților din probe, se numesc **date** sau **observații**.

În ecologie, de multe ori, se **numără entitățile** dintr-un grup sau dintr-o colecție. Pentru ca o astfel de numărătoare să aibă valoare, trebuie specificată dimensiunea grupului sau colecției, care se numește **unitatea de probă**. Un set de unități de probă alcătuiesc o probă, iar observația este numărul de entități dintr-o anumită unitate de probă.

Principala diferență dintre numărători și măsurători este, în cazul măsurătorilor, lipsa unui control asupra dimensiunii unităților de probă. Atunci când se numără entități, se poate decide care este unitatea de probă. Conținutul unei capcane sau pătrat de probă reprezintă o probă dacă entitățile investigate sunt măsurate, și o unitate de probă dacă entitățile sunt doar numărate.

O întrebare cu un răspuns nu întotdeauna evident se referă la identificarea populației din care provin unitățile de probă. Dacă ceea ce s-a capturat din zece capcane de sol constituie o probă, care este populația din care a fost extrasă această probă? În acest caz, populația este reprezentată de numărul total de capcane care ar fi putut fi instalate în întreaga suprafață de studiu. O astfel de populație este una ipotetică.

Pentru ca o probă să fie reprezentativă pentru populația din care a fost extrasă este necesar ca prelevarea acesteia să fie făcută **aleator, randomizat** sau **la întâmplare**. Aceasta înseamnă că unicul criteriu folosit în extragerea unităților de probă este ca toate unitățile să aibă șanse egale de a face parte din probă. De exemplu, dacă proba se obține cu ajutorul unor capcane cu momeală, animalele care vor cădea în acestea vor fi cele mai flămânde și eventual cu o greutate mai mică. Astfel, proba obținută nu va fi reprezentativă pentru populație, deoarece animalele cu o greutate mai mică au șanse mai mari să fie capturate decât cele cu o greutate mai mare și care sunt mai sătule. Dacă proba nu este reprezentativă, atunci generalizările care se vor face pornind de la aceasta, cu privire la întreaga populație, vor fi eronate. Dacă ne referim la exemplul anterior, proba obținută cu ajutorul

capcanelor cu momeală este reprezentativă pentru populația statistică a animalelor flămânde din populația biologică. Obținerea unor probe aleatoare este asigurată de metodele de lucru utilizate în funcție de caracteristicile entităților urmărite. În cazul în care există suspiciunea că o probă nu este aleatoare, acest lucru trebuie specificat sau concluziile rezultate prin extrapolare trebuie legate de populația statistică din care proba a fost extrasă.

Un alt aspect important este reprezentat de **independența** observațiilor din probe. Aceasta se referă la faptul că apariția unei anumite valori individuale într-o probă nu influențează probabilitatea de apariție în probă a altei valori. De exemplu, dacă se studiază o populație ipotetică formată din zece entități, probabilitatea de a extrage o entitate este de 1/10, iar dacă nu se face reintroducere, probabilitatea de a extrage următoarea entitate este de 1/9 și așa mai departe. Deci extragerea unei entități modifică probabilitatea de extragere a celorlalte și observațiile nu sunt independente. O astfel de situație nu trebuie să constituie un motiv de preocupare în cazul populațiilor mari, așa cum sunt majoritatea populațiilor biologice.

Uneori, obținerea unor observații neindependente este intenționată. De exemplu, când se dorește studierea efectului unui anumit tratament asupra variabilei, se fac observații repetate asupra aceloraşi entități pentru a evidenția dacă deosebirile dintre observații sunt diferite semnificativ, adică dacă tratamentul a modificat valorile variabilei semnificativ. Dacă însă se fac observații repetate asupra unei singure entități, atunci concluziile rezultate nu pot fi extrapolate la populația de proveniență a entității respective, deoarece proba cu dimensiunea unu nu este reprezentativă.

Un descriptor sau o măsură a unei variabile în probă se numește **statistică**. O statistică a unei probe se folosește de obicei pentru a estima un **parametru** al populației. De exemplu, media valorilor dintr-o probă este o statistică, iar media populației din care a fost extrasă proba, un parametru. Dat fiind faptul că în ecologie sunt rare cazurile în care se poate afla media unei populații prin investigarea fiecărei unități de probă, media populației respective poate fi estimată pornind de la statistica probei reprezentative.

Populațiile ipotetice au parametri ipotetici și sunt de obicei folosite pentru comparații. De exemplu, media numărului de plante dintr-o anumită specie din zece pătrate de 1 m^2 este o estimare a mediei numărului de plante

per pătrat sau metru pătrat, adică parametrul populației de pătrate care s-ar putea delimita în aria de studiu. Astfel de parametri sunt utili atunci când se compară diferite zone de studiu (două zone stepice, două zone de pădure, două bazine acvatice etc.).

Atunci când se estimează un parametru populațional pornind de la statistica corespunzătoare, dimensiunea probei sau numărul unităților de probă are o mare importanță. În general, cu cât proba are o dimensiune mai mare, cu atât va fi mai reprezentativă pentru populația de proveniență și estimarea parametrilor mai precisă. Totuși obținerea unor probe extrem de numeroase este uneori imposibilă sau presupune un efort foarte mare care ar putea fi investit în alte direcții de cercetare. Astfel, este bine să existe un echilibru între aceste două aspecte, între dimensiunea probei și efortul necesar obținerii acesteia.

2. APRECIEREA ŞI PREZENTAREA DATELOR

Variabilele sunt caracteristici sau caractere care au valori ce pot fi diferite de la un individ la altul, într-o populaţie. Deci o variabilă poate lua mai multe **valori individuale** în populaţia studiată. Valorile individuale determinate prin investigarea unor indivizi sau unităţi dintr-o probă se numesc **date**.

2.1. SCALE DE MĂSURARE ŞI TIPURI DE VARIABILE

De exemplu, într-o populaţie de peşti dintr-o anumită specie, se investighează lungimea indivizilor. Lungimea reprezintă o caracteristică sau un caracter al tuturor indivizilor din populaţie, deci este o variabilă. Lungimea peştilor are valori diferite de la un individ la altul sau de la un grup de indivizi la altul. Aceste valori ale tuturor indivizilor din populaţie se numesc valori individuale. Dacă se capturează un anumit număr de peşti (se extrage o probă) din populaţia de studiu şi se măsoară lungimea fiecărui individ, valorile individuale astfel determinate constituie datele.

În funcţie de relaţiile ce se pot stabili între valorile individuale ale unei variabile, aceasta poate aparţine unui anumit tip de variabilă, care la rândul său poate fi apreciată pe o anumită scală cu anumite reguli şi limitări. În general, sunt recunoscute patru astfel de scale de apreciere a variabilelor: **nominală**, **ordinală**, **de interval** şi **de raport**. Relaţia dintre aceste scale este una ierarhică, adică o scală de nivel superior încorporează proprietăţile scalelor inferioare acesteia.

Scala nominală. Este cea mai simplă modalitate de apreciere a variabilelor. În esenţă, permite doar identificarea categoriilor în care valorile individuale pot fi clasificate. Categoriile se exclud reciproc, adică o anumită valoare poate aparţine doar unei singure categorii din scală. Variabilele corespunzătoare acestei scale se numesc **variabile nominale** sau **atribute**.

Oricare două valori ale unei astfel de variabile pot aparţine aceleiaşi categorii sau la două categorii diferite ale scalei ordinale, cu alte cuvinte pot fi egale sau diferite (tab. 2.2).

De exemplu, sexul indivizilor unei populaţii este o variabilă nominală, ale cărei valori individuale posibile sunt mascul şi femel. Doi indivizi dintr-o populaţie pot avea acelaşi sex (mascul sau femel) sau pot avea sexe diferite (unul este mascul, celălalt este femel) la un moment dat. Deci valorile variabilei sex pot fi egale (aparţin aceleiaşi categorii a scalei nominale) sau diferite (aparţin la categorii diferite ale scalei nominale). Alte exemple de variabile nominale întâlnite în ecologie sunt: culoarea, tipul de habitat, specia.

Scala ordinală. Aceasta include proprietăţile scalei nominale (identificare şi clasificare), la care se mai adaugă posibilitatea de ordonare a categoriilor într-o serie, de la valoarea cea mai mică la cea mai mare, sau de specificare a magnitudinii acestora. Variabilele corespunzătoare acestei scale se numesc **variabile ordinale**. Oricare două valori ale unei astfel de variabile pot fi egale sau diferite. În cazul în care sunt diferite, valorile se pot ordona, adică se poate spune că una dintre valori este mai mare decât cealaltă (tab. 2.2). În general, valorile variabilelor ordinale se reprezintă sub formă de magnitudine relativă.

De exemplu, dacă într-o populaţie de lupi urmărim variabila agresivitate, valorile individuale pot fi „neagresiv", „puţin agresiv", „agresiv" şi „foarte agresiv". Doi indivizi pot fi egali sau diferiţi din punctul de vedere al agresivităţii, iar dacă sunt diferiţi, atunci se poate determina că unul este mai agresiv decât celălalt (că o valoare este mai mare decât cealaltă), dar nu se poate spune exact cu cât.

O altă modalitate de a reprezenta valorile pe scala ordinală constă în folosirea unor simboluri numerice corespunzătoare magnitudinii valorilor, numite **ranguri**. Rangurile sunt utile mai ales când se urmăreşte reprezentarea unei variabile pe o scală cu mai multe categorii ordinale. Astfel, valoarea cea mai mică, „neagresiv" din exemplul anterior, primeşte rangul unu, următoarea doi şi aşa mai departe.

Un exemplu de scală ordinală este scala DAFOR (acronimul este format prin preluarea primei litere a valorilor scalei), utilizată pentru aprecierea abundenţei unei specii de plante într-un pătrat de probă (tab. 2.1).

Tabelul 2.1

Valoare	Dominată	Abundentă	Frecventă	Ocazională	Rară
Rang	5	4	3	2	1

Trebuie reținut că valorile numerice ale rangurilor nu pot fi folosite pentru a efectua operații simple, deoarece acestea nu au sens – valoarea „abundentă" nu este de două ori mai mare decât valoarea „ocazională" sau diferența dintre valoare „dominantă" și valoarea „abundentă" poate să nu fie egală cu cea dintre valorile „ocazională" și „rară". Rangurile sunt doar simboluri numerice care arată magnitudinea valorilor sau poziția lor în setul de valori ordonate.

Scala de interval. Permite atât ordonarea datelor, cât și precizarea distanței dintre unitățile scalei. Valorile exprimate pe această scală pot fi scăzute unele din altele pentru a afla exact care este diferența dintre ele. Din cauza faptului că scala de interval nu are o valoare zero absolută, nu se poate realiza împărțirea valorilor pentru a afla cu cât una este mai mare decât cealaltă (tab. 2.2).

De exemplu, variabila de tip dată este apreciată pe o scală de interval. Dacă trei specii de păsări de talie mică (paseriforme) revin din migrație pe 5, 10 și 15 mai, putem spune că a treia specie ajunge cu zece zile mai târziu decât prima, dar nu putem spune că are nevoie de trei ori mai mult timp pentru a încheia migrația. Un alt exemplu de scală de interval este scala Celsius de apreciere a temperaturii: $0°C$ este o valoare convențională, aleasă să desemneze temperatura de îngheț a apei. Ca urmare, o temperatură de $10°C$ nu înseamnă de două ori mai cald decât $5°C$. Datorită faptului că scala Celsius de apreciere a temperaturii nu are un zero absolut, aceasta prezintă și valori negative care nu ar putea exista în cazul unei valori zero absolute.

Scala de raport. Pe lângă proprietățile celorlalte scale (identificare, clasificare, ordonare, precizarea diferenței), aceasta mai permite și împărțirea valorilor unele la altele pentru a putea afla de câte ori una este mai mare decât cealaltă (tab. 2.2).

De exemplu, lungimea se apreciază pe scală de raport, iar o lungime de 30 cm este de trei ori mai mare decât una de 10 cm.

Această proprietate este posibilă datorită faptului că scalele de raport au valori zero absolute, adică zero înseamnă nimic, vid. Ca urmare, aceste scale nu pot prezenta valori negative.

De exemplu, scala Kelvin de apreciere a temperaturii este o scală de raport, deoarece valoarea $0K$ $(-273{,}15°C)$ reprezintă temperatura față de care nimic nu poate fi mai rece și la care în materie nu mai există energie sub formă de căldură.

Variabilele corespunzătoare scalei de interval și scalei de raport pot fi de două tipuri: **discontinue** și **continue**.

Variabilele discontinue sau **discrete** pot lua anumite valori (de obicei, întregi și pozitive), între care nu există valori intermediare. Aceste variabile reprezintă caractere numărabile sau meristice (număr de solzi, număr de ouă, număr de elemente florale, număr de pui etc.). De exemplu, dimensiunea pontei unei păsări este o variabilă discretă, ale cărei valori sunt întregi și pozitive; nu există cuiburi cu număr fracționar de ouă.

Variabilele continue pot lua orice valoare dintr-un anumit interval, iar între oricare două valori există o infinitate de valori posibile. Aceste variabile reprezintă caractere măsurabile sau metrice (lungime, lățime, înălțime, greutate, temperatură etc.). De exemplu, între $10\ cm$ și $11\ cm$ pot exista, în principiu, o infinitate de valori posibile, în funcție de numărul zecimalelor considerate.

Tabelul 2.2. Caracteristicile esențiale ale tipurilor de variabile

Scala de apreciere a variabilelor	Semnele care se pot pune între valori
Nominală	=; ≠
Ordinală	=; ≠; <; >
De interval	=; ≠; <; >; -
De raport	=; ≠; <; >; -; :

Conversia datelor de la o scală la alta

Conversia datelor se poate face doar în sensul pierderii unei părți din informația deținută de acestea. Ca urmare, conversia se poate face doar de la o scală superioară ierarhic către una de nivel inferior. Astfel, datele măsurate pe o scală de interval sau de raport pot fi convertite la o scală ordinală sau nominală. Datele măsurate pe o scală ordinală pot fi convertite doar la o scală nominală.

Exemplul 2.1. Dacă s-au determinat înălțimile unor plante de stepă în centimetri, variabila urmărită, înălțimea, este una continuă, exprimată pe o scală de raport. Pentru a realiza conversia la o scală ordinală, se dau ranguri valorilor inițiale. Astfel, valoarea cea mai mică va primi rangul 1, următoarea 2, iar valoarea cea mai mare va primi rangul maxim. Înălțimea de 13 cm va primi rangul 1, ceea ce arată că este vorba de valoarea cea mai mică, iar înălțimea de 23 cm va primi rangul 9, adică valoarea maximă din probă. Valorile egale ale înălțimii vor primi media rangurilor pe care le-ar fi primit dacă ar fi fost diferite. Observăm în tabelul 2.3 că există trei valori egale, de 15 cm, și alte două valori egale între ele, de 17 cm. Cele trei valori de 15 cm, dacă ar fi fost diferite, ar fi primit rangurile 2, 3 și 4. Fiind egale, primesc media rangurilor pe care le-ar fi primit dacă ar fi fost diferite, adică $(2 + 3 + 4)/3 = 3$. La fel se procedează și în cazul celor două valori de 17 cm – media rangurilor pe care le-ar fi primit dacă ar fi fost diferite este $(5 + 6)/2 = 5{,}5$. În continuare, pentru conversia la o scală ordinală, se consideră o valoare de referință a înălțimii din probă, după care toate celelalte valori se exprimă în relație cu aceasta: egale cu valoarea de referință sau diferite de aceasta. Dacă din anumite motive ne interesează plantele cu înălțimea de 17 cm, atunci vom avea două valori „= 17" și șapte valori „≠ 17"

Tabelul 2.3

Înălțimea (cm), scală de interval sau raport	13	15	15	15	17	17	19	21	23
Ranguri intermediare (dacă valorile ar fi diferite)	1	2	3	4	5	6	7	8	9
Ranguri, scală ordinală	1	3	3	3	5,5	5,5	7	8	9
Valori nominale, scală nominală	≠17	≠17	≠17	≠17	=17	=17	≠17	≠17	≠17

Cea mai frecventă conversie este cea de la datele apreciate pe o scală de interval sau de raport, la una ordinală. O astfel de conversie este necesară atunci când datele vor fi analizate prin metode neparametrice, deoarece nu sunt îndeplinite condițiile de aplicare ale metodelor parametrice.

Variabile derivate

În anumite situații, variabilele originale sunt procesate matematic,

astfel încât să rezulte variabile derivate cum ar fi: **rapoarte, proporții, procente** și **rate**.

Raportul este o relație simplă între două numere. De exemplu, dacă lungimea capului la o viperă de stepă este 17,7 mm și lățimea de 11,7 mm, raportul lungime:lățime este de 17,7: 11,7. Implicit, raportul lățime:lungime este de 11,7: 17,7. Uneori, una dintre valori poate fi convertită prin împărțire la unitate. De exemplu, dacă într-o probă sunt 19 $masculi$ și 27 $femele$, atunci raportul masculi:femele este 19: 27 sau 1: 27/19, adică 1: 1,421. Raportul poate fi scris și ca o fracție. În cazul exemplului anterior, raportul dintre masculi și femele este de $19/27 = 1/1{,}421$. Rezultatul calculării fracției se numește **coeficient**; astfel, $1/1{,}421 = 0{,}704$.

Proporția este raportul dintre parte și întreg. Dacă lungimea totală a unei vipere de stepă este 490 mm, iar lungimea cozii este 65 mm, proporția reprezentată de coadă este $65: 490 = 0{,}13$. Dacă se calculează o proporție pornind de la raportul dintre numărul de valori dintr-o categorie și numărul total de valori din toate categoriile, atunci aceasta se numește **frecvență proporțională**.

Procentul se obține prin înmulțirea valorii unei proporții cu 100.

Rata reprezintă raportarea unei observații la unitatea de timp. Ratele se folosesc pentru a exprima anumite variabile cum ar fi creșterea, dinamica unei populații, mișcarea.

De exemplu, dacă o plantulă crește 15 cm în 10 $zile$, atunci rata de creștere este de $15/10 = 1{,}5\ cm/zi$.

Numeroși indici ecologici cum ar fi indicii de diversitate sunt de fapt variabile derivate. Uneori acestea pot fi analizate prin metode statistice dar numai după o conversie sau transformare prealabilă a datelor.

2.2. REPREZENTAREA DATELOR

Unul dintre inconvenientele majore ale prezentării datelor sub formă de tabele constă în faptul că informația nu este evidentă imediat. Ea poate fi percepută doar după o analiză în detaliu a fiecărei valori sau a majorității valorilor din tabel. Pentru facilitarea percepției informației conținute de date, este necesară procesarea și transformarea acestora într-o prezentare vizuală. Modalitatea cea mai des utilizată de prezentare a datelor constă în

folosirea reprezentărilor grafice. Tipul de reprezentare grafică se alege în funcție de tipul de variabilă.

Reprezentarea variabilelor discrete. Procesarea datelor constă în acest caz în aranjarea lor în tabelul de distribuție a frecvențelor, adică se prezintă fiecare valoare a variabilei și frecvența corespunzătoare acesteia, adică de câte ori se întâlnește o anumită valoare în probă.

Exemplul 2.2. Într-un studiu se urmărește numărul de fitoindivizi (de plante) din specia *Crambe tataria* în 20 de pătrate de 10x10 m dintr-o pajiște stepică. Tabelul de distribuție a frecvențelor se prezintă astfel:

Tabelul 2.4

Nr. fitoindivizi/pătrat (x)	0	1	2	5	7	10	16	19	38	60
Frecvența (f)	5	4	3	1	1	2	1	1	1	1

În continuare se reprezintă grafic pe abscisă valorile ordonate ale variabilei (x), iar pe ordonată valorile frecvențelor (f) corespunzătoare valorilor variabilei. Practic, frecvența fiecărei valori a variabilei este reprezentată printr-o coloană cu înălțime corespunzătoare. Se obține astfel o diagramă în coloane (dreptunghiuri) a distribuției frecvențelor unei variabile discrete. Trebuie remarcat spațiul dintre coloanele corespunzătoare valorilor ordonate ale variabilei – acesta sugerează absența valorilor intermediare dintre valorile alăturate ale unei variabile discrete (fig. 2.1).

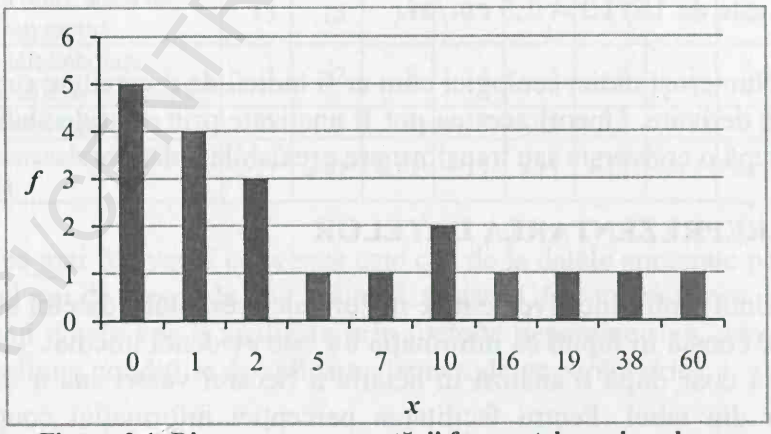


Figura 2.1. Diagrama reprezentării frecvențelor prin coloane

Diagrama poate fi realizată și prin reprezentarea frecvențelor prin

puncte (fig. 2.2). În acest caz, se impune ca acestea să nu fie unite, pentru a sugera, ca şi în cazul spaţiului dintre coloanele graficului anterior, că este vorba de valorile unei variabile discontinue.

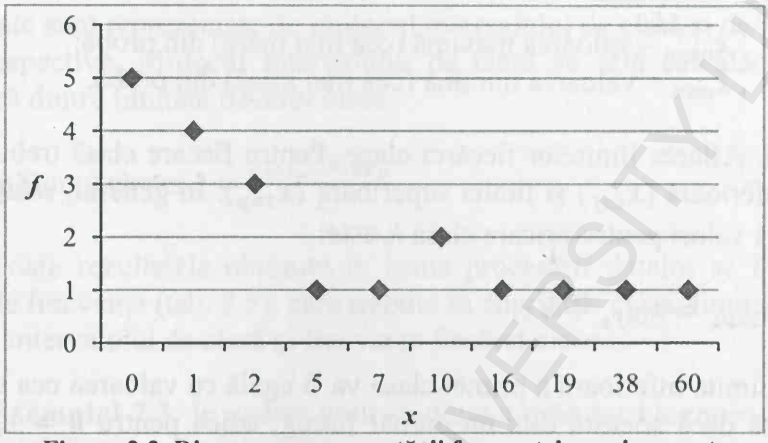


Figura 2.2. Diagrama reprezentării frecvenţelor prin puncte

Aceleaşi tipuri de reprezentări pot fi utilizate şi pentru reprezentarea distribuţiei frecvenţelor **variabilelor nominale** şi a celor **ordinale**. În cazul variabilelor nominale, ordinea valorilor acestora pe abscisă este arbitrară.

Reprezentarea variabilelor continue. Datorită faptului că variabilele continue iau valori din aproape în aproape, există posibilitatea ca o probă să nu conţină nici măcar două valori identice. Ca urmare, nu se mai poate opera cu frecvenţa unei singure valori, pentru că, într-o astfel de situaţie, toate valorile fiind diferite, vor avea frecvenţa egală cu 1, adică vor apărea în probă o singură dată. Astfel, în cazul în care se doreşte reprezentarea distribuţiei frecvenţelor valorilor unei variabile continue, este necesară gruparea valorilor din probă în clase de frecvenţă, ceea ce implică parcurgerea mai multor etape de procesare a datelor:

1. Aflarea numărului de clase. Numărul de clase (k) este rezultatul rotunjit la cel mai apropiat întreg, ce se poate afla folosind una din următoarele două relaţii:

$$k = 1 + 3{,}3 \cdot log_{10}(n) \text{ sau } k < 5 \cdot log_{10}(n)$$

n – numărul de valori din probă.

2. Aflarea intervalului de clasă. Intervalul de clasă (h) este rezultatul relației:

$$h = \frac{x_{max} - x_{min}}{k}$$

x_{max} – valoarea maximă (cea mai mare) din probă;
x_{min} – valoarea minimă (cea mai mică) din probă.

3. Aflarea limitelor fiecărei clase. Pentru fiecare clasă trebuie aflată limita inferioară (x_{inf}) și limita superioară (x_{sup}). În general, relația dintre cele două valori pentru oricare clasă k este:

$$x_{sup_k} = x_{inf_k} + h\,.$$

Limita inferioară a primei clase va fi egală cu valoarea cea mai mică din probă dacă aceasta este un număr întreg, adică pentru $k = 1$, $x_{inf_1} = x_{min}$. Dacă x_{min} nu este un întreg, atunci x_{inf_1} va fi întregul aflat prin rotunjirea prin lipsă al lui x_{min}. Limita superioară a primei clase se va afla adunând la valoarea minimă sau la limita inferioară a clasei valoarea intervalului de clasă, conform relației:

$$x_{sup_1} = x_{inf_1} + h \text{ sau } x_{sup_1} = x_{min} + h\,.$$

Pentru a afla x_{inf_2}, la x_{sup_1} se va adăuga 1. Astfel, între cele două clase nu va exista nici un fel de suprapunere, adică o valoare din probă egală cu x_{sup_1} va face parte doar din prima clasă. În acest fel se vor afla limitele celorlalte clase. Ultima clasă, k, va trebui să includă valoarea cea mai mare din probă, adică pe x_{max}.

4. Aflarea frecvenței fiecărei clase. Frecvența claselor se va afla prin numărarea valorilor din probă cuprinse între limita inferioară și cea superioară a fiecărei clase. În final, trebuie ca fiecare valoare din probă să fie inclusă într-una din clase. Suma frecvențelor tuturor claselor trebuie să fie egală cu numărul de valori din probe, adică $\sum f = n$.

Frecvențele claselor se pot reprezenta sub forma unei histograme. Spre deosebire de diagrama frecvențelor prezentată pentru variabilele

discrete, histograma are coloanele unite, ceea ce sugerează continuitatea dintre clasele de frecvență ale valorilor unei variabile continue.

O altă modalitate de reprezentare a frecvențelor claselor este și poligonul frecvențelor. Acesta se construiește prin unirea punctelor ale căror coordonate sunt reprezentate de mijlocul intervalului de clasă și de frecvența clasei respective. Mijlocul intervalului de clasă se află calculând media aritmetică dintre limitele fiecărei clase:

$$\text{mijlocul clasei } k = \frac{x_{inf_k} + x_{sup_k}}{2}.$$

Toate rezultatele obținute în urma procesării datelor se înscriu în tabelul de frecvențe (tab. 2.5), care trebuie să cuprindă: clasa, limitele clasei, mijlocul intervalului de clasă și frecvența fiecărei clase.

Exemplul 2.3. În cadrul unui studiu s-a măsurat lungimea în mm a 100 de pești dintr-o anumită specie.

194	140	226	269	284	243	303	235	229	239
206	262	233	307	285	180	248	205	284	191
154	224	307	236	198	288	241	252	385	220
299	273	275	164	137	357	246	271	246	276
229	280	227	253	286	190	291	297	296	288
225	234	244	351	267	265	239	283	190	244
288	245	289	241	289	278	255	253	240	153
208	328	235	283	214	300	228	204	343	228
194	233	218	321	303	254	225	232	196	245
223	305	220	338	269	224	319	259	240	293

$n = 100$

$k = 1 + 3{,}3 \cdot log_{10}(100) = 7{,}6 \approx 8$

$x_{max} = 385$

$x_{min} = 137$

$h = \frac{385-137}{8} = 31$

Tabelul 2.5. Tabelul de distribuție a frecvențelor claselor de lungime (mm)

k	x_{inf}	x_{sup}	mijloc	f	F (f cumulată)
1	137	168	152,5	5	5
2	168	199	183,5	8	13
3	199	230	214,5	19	32
4	230	261	245,5	27	59
5	261	292	276,5	23	82
6	292	323	307,5	12	94
7	323	354	338,5	4	98
8	354	385	369,5	2	100

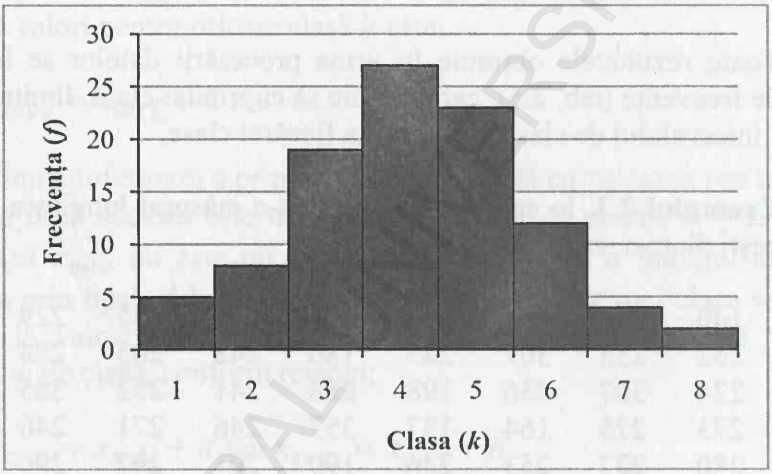


Figura 2.3. Histograma frecvențelor

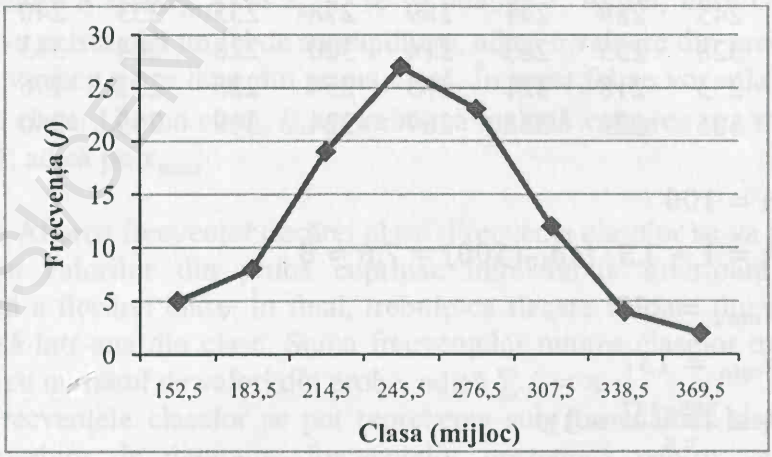


Figura 2.4. Poligonul frecvențelor

3. DESCRIEREA STATISTICĂ A PROBELOR ECOLOGICE

Statistica descriptivă este partea statisticii care se ocupă de culegerea și de clasificarea datelor statistice și, pe această bază, de descrierea fenomenelor investigate. Rolul ei este de a rezuma cantitativ informația culeasă, de a descrie și de a pune în evidență esențialul, în fine, de a realiza sinteze cu ajutorul unui limbaj numeric.

În natură, atunci când investigăm o populație, rareori întâlnim valori individuale identice ale unor variabile. La o privire mai atentă a datelor, se poate observa existența unor valori în jurul cărora tind să se distribuie majoritatea, dacă nu toate celelalte valori individuale. Descrierea statistică a probelor prelevate din populații scoate în evidență două aspecte esențiale: **tendința centrală** și **variabilitatea** valorilor individuale.

3.1. TENDINȚA CENTRALĂ

Tendința centrală a unor date reprezintă o valoare sau o condiție reprezentativă pentru toate datele din probă sau pentru valorile individuale din populație. De exemplu, enunțuri ca „majoritatea florilor dintr-o probă au culoarea roșie" sau „diametrul mediu al florilor este de 2 cm" surprind tocmai această tendință centrală a valorilor individuale din probe.

În funcție de scala de apreciere a datelor din probă și implicit a caracteristicilor tipului de variabilă urmărită, există mai multe măsuri sau descriptori ai tendinței centrale, dintre care cei mai des utilizați sunt: **modul**, **mediana** și **media**.

Modul (*Mo*). Măsura tendinței centrale, reprezentată de valoarea din probă cu frecvența cea mai mare, adică cel mai des întâlnită, se numește **mod**. De exemplu, dacă o probă este reprezentată de 20 de plante la care se urmărește culoarea florilor, iar zece dintre acestea au flori de culoare roșie, șapte au flori de culoare violet și trei sunt grena, atunci modul probei va fi

valoarea „roşu" a variabilei „culoarea florilor". Aşa cum reiese din acest exemplu, modul se poate folosi pentru descrierea tendinţei centrale a unei variabile apreciate pe o scală nominală; de fapt este singurul descriptor de acest fel ce se poate folosi pentru valorile unei variabile nominale. Modul se poate folosi şi pentru celelalte tipuri de variabile – ordinale, discrete şi continue. În cazul variabilelor continue, se poate ca modul să nu poată fi aplicat. Dat fiind faptul că aceste variabile iau valori din aproape în aproape şi că între oricare valori alăturate există un număr infinit de valori posibile, se poate întâmpla ca într-o probă toate valorile să fie diferite, caz în care frecvenţa fiecărei valori va fi egală cu unu. Deci, într-o astfel de situaţie, nu există nici o valoare cu frecvenţă mai mare decât celelalte şi modul nu se poate calcula decât pentru clasele de frecvenţă.

În exemplul 2.3 lungimea peştilor este o variabilă continuă. Modul probei este 288, pentru care există 3 valori. Dacă toate valorile ar fi fost diferite, atunci s-ar fi putut afla **clasa modală** de distribuţie a frecvenţelor ca fiind clasa cu frecvenţa cea mai mare. În cazul exemplului luat în discuţie, clasa modală este clasa nr. 4 cu frecvenţa 27.

Dacă într-o distribuţie a frecvenţelor apar mai multe vârfuri de frecvenţă sau moduri, aceasta va fi numită **distribuţie multimodală**. Multe caractere răspunzătoare de dimorfismul sexual prezintă în populaţii o distribuţie **bimodală**. În figura 3.1 apare o astfel de distribuţie bimodală: modul pentru femele este 29 de plăci subcaudale, iar pentru masculi este 37.

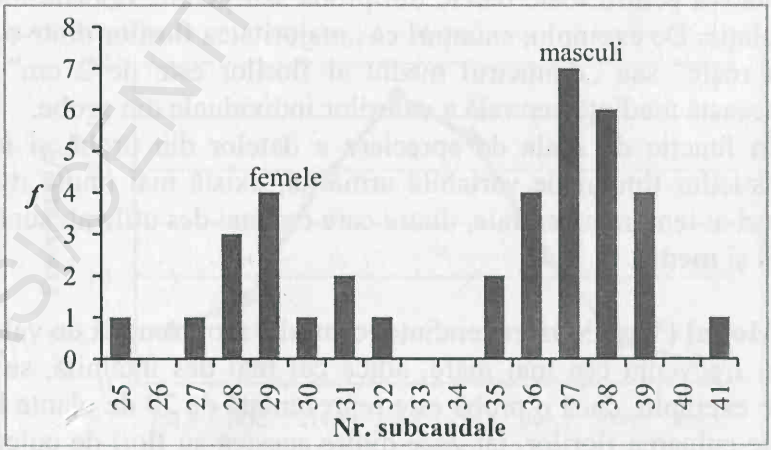


Figura 3.1. Diagrama distribuţiei frecvenţelor numărului de subcaudale la *Vipera ursinii moldavica*

În cazul unor astfel de distribuții, care nu sunt simetrice față de o singură valoare a tendinței centrale, se recomandă ca analiza statistică să se facă separat, pe categorii discrete din probă – masculi, femele, juvenili – pentru care datele se prezintă mai mult sau mai puțin simetrice.

Mediana (*Me*). Este măsura tendinței centrale care reprezintă valoarea centrală sau media valorilor centrale ale unui set de date ordonate crescător. Valoarea centrală este cea care, în setul de date ordonat crescător, este precedată și succedată de același număr de valori individuale. Rezultă că mediana se poate utiliza pentru date care se pot ordona, adică pentru date măsurate pe o scală de raport, interval sau ordinală, și se poate folosi pentru datele apreciate pe o scală nominală. Este considerată una dintre cele mai robuste măsuri ale tendinței centrale, deoarece nu este influențată de eventualele valori atipice, cum se poate întâmpla în cazul mediei, și se poate utiliza chiar și în cazurile în care se cunosc doar magnitudinile (sau rangurile) unor valori ce nu au fost înregistrate.

Modalitatea de calcul a medianei depinde de numărul valorilor din probă (n):

– dacă proba are un număr par de date ($n = 2k + 1$), atunci mediana va fi reprezentată de valoarea centrală:

pt. $n = 2k + 1$, atunci $Me = x_{k+1}$ sau $Me = x_{(n+1)/2}$.

De exemplu, pentru seria de date 1, 2, 2, 3, 4, 4, 5, $Me = 3$, pentru că are un număr egal de valori la stânga și la dreapta sa.

– dacă proba are un număr par de date ($n = 2k$), atunci mediana va fi reprezentată de media celor două valori centrale sau de intervalul median:

pt. $n = 2k$, atunci $Me = \frac{x_k + x_{k+1}}{2}$ sau $Me = \frac{x_{n/2} + x_{(n/2)+1}}{2}$.

De exemplu, pentru seria de date 1, 2, 2, 3, 4, 5, 6, 7, cele două valori centrale sunt 3 și 4, deci $Me = (3 + 4)/2 = 3{,}5$.

În cazul în care există numeroase observații cu aceeași valoare cu cea a medianei datorită prezentării datelor sub formă de intervale de clasă,

atunci formulele de calcul se modifică astfel:

pt. $n = 2k + 1$, atunci $Me = x_{inf Me} + h \cdot \frac{\frac{1}{2}\cdot(\sum f) - F_{Me-1}}{f_{Me}}$;

pt. $n = 2k$, atunci $Me = x_{inf Me} + h \cdot \frac{\frac{1}{2}\cdot(1+\sum f) - F_{Me-1}}{f_{Me}}$,

$x_{inf\,Me}$ – limita inferioară a clasei mediane (clasa de frecvență ce conține valoarea mediană);

h – valoarea intervalului de clasă;

$\sum f$ – suma frecvențelor tuturor claselor;

F_{Me-1} – frecvența cumulată a clasei dinaintea clasei mediane (suma frecvențelor claselor care preced clasa mediană);

f_{Me} – frecvența clasei mediane.

Exemplu 3.1. Într-un studiu s-a urmărit acoperirea procentuală realizată de o specie de ierburi de stepă, în 20 de suprafețe de probă.

Acoperire	80-100%	60-80%	40-60%	20-40%	0-20%	$\sum f$
Frecvența (f)	1	2	9	6	2	20

Conform definiției medianei, aceasta ar trebuie să reprezinte media acoperirilor din suprafețele de probă 10 și 11. Valorile exacte ale acoperirii nu sunt accesibile, acestea fiind aproximate prin clase de acoperire. Acestea sunt echivalente din punct de vedere statistic unor clase de distribuție a frecvențelor. Dacă simbolizăm fiecare observație prin procentul mediu al fiecărei clase, datele se prezintă astfel:

10	10	30	30	30	30	30	30	50	**50**
50	50	50	50	50	50	50	70	70	90

Cifrele scrise îngroșat ar reprezenta cele două valori centrale necesare calculării medianei în cazul unui număr par de date. Se observă că sunt mai multe valori egale cu 50; sunt 9 valori în clasa (intervalul) mediană. Frecvența cumulată a clasei de dinaintea celei mediane este $2 + 6 = 8$. Deci, până la prima valoare centrală, mai sunt 2 suprafețe de probă. Rezultă că prima valoare centrală se găsește la 2/9 din intervalul clasei mediane. Intervalul de clasă este egal cu 20%; 2/9 din 20 reprezintă

4,45. Dacă la această valoare adăugăm limita inferioară a clasei mediane (40), obținem 44,45, adică prima valoare centrală a acoperirii. Folosind același raționament, se obține și a doua valoare centrală: $40 + 20 \cdot 3/9 = 46{,}67$. Mediana va fi egală cu media celor două valori centrale: $(44{,}45 + 46{,}67)/2 = 45{,}56$.

O altă modalitate constă în aplicarea formulei de mai sus, care are la bază același raționament:

$x_{inf\,Me} = 40$
$h = 20$
$\sum f = 20$
$F_{Me-1} = 8$
$f_{Me} = 9$

$$Me = x_{inf Me} + h \cdot \frac{\frac{1}{2} \cdot (1 + \sum f) - F_{Me-1}}{f_{Me}}$$

$$Me = 40 + 20 \cdot \frac{\frac{1}{2} \cdot (1 + 20) - 8}{9} = 40 + 20 \cdot \frac{2{,}5}{9} = 45{,}56\,.$$

Media ($\bar{x}, \mu$). Este unul dintre cei mai cunoscuți și mai utili descriptori ai tendinței centrale. Există mai multe tipuri de medie, dar cel mai utilizat este media aritmetică. Dacă se calculează media luând în considerație toți indivizii unei populații, atunci aceasta se numește **medie populațională**, este un parametru populațional și se notează cu μ. Media obținută în urma analizei datelor dintr-o probă sau **media probei** este o statistică a probei simbolizată cu $\bar{x}$. Media probei poate fi un estimator direct al mediei populaționale ($\bar{x}$ estimează pe μ).

Formula de calcul a mediei este suma tuturor valorilor supra numărul acestora:

– pentru populație $\mu = \frac{\sum x}{N}$;

– pentru probă $\bar{x} = \frac{\sum x}{n}$,

x – fiecare valoare individuală din populație sau probă;

N – numărul tuturor valorilor individuale din populație;

n – numărul tuturor valorilor din proba prelevată din populație.

Relația dintre medie, mediană și mod.

În cazul unei variabile a cărei valori individuale sunt distribuite perfect simetric, media, mediana și modul sunt egale. În cazul unei

distribuții ușor asimetrice unimodale, mediana se dispune între medie și mod. În majoritatea distribuțiilor biologice se observă o abatere pozitivă, adică media are valoare mai mare decât mediana, care la rândul ei este mai mare decât modul (fig. 3.2).

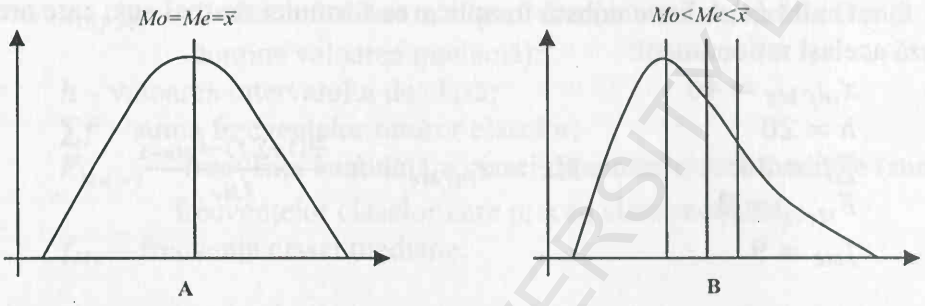


Figura 3.2. Relația dintre mod, mediană și medie: A – distribuție simetrică; B – distribuție asimetrică

Dintre cele trei măsuri ale tendinței centrale, media este singura care ține cont de toate datele din probe și astfel sintetizează întreaga informație furnizată de acestea. Valoarea mediei este folosită în numeroase tehnici de analiză statistică. De asemenea, poate fi combinată cu mediile altor probe din aceeași populație, în cadrul mediei generale, atunci când datele sunt rare sau greu de obținut.

$$\text{media generală} = \frac{\sum(n_i \cdot \bar{x}_i)}{\sum n_i}$$

$\bar{x}_i$ – media probei i $\quad$ n_i – numărul de valori din proba i

Media este însă ușor influențată de apariția unor valori atipice (foarte mari sau foarte mici față de majoritatea valorilor). În astfel de situații, mediana reprezintă un descriptor mai robust al tendinței centrale a probei. Mediana este utilă și în analiza preliminară, deoarece scoate în evidență tendințele generale ale datelor.

Modul reprezintă o modalitate rapidă și aproximativă de a aprecia tendința centrală a probei și de a indica centrul distribuției observațiilor, apreciate pe o scală ordinală sau nominală.

3.2. VARIABILITATEA

Variabilitatea este o trăsătură generală a sistemelor naturale. Este foarte puțin probabil ca indivizii dintr-o populație biologică să fie identici din punctul de vedere al unui caracter sau al unei variabile. Dacă nu ar exista variabilitate, nu ar mai fi nevoie de analiza statistică – o singură valoare ar fi suficientă pentru a descrie variabila cercetată pentru întreaga populație (implicit, nici tendința centrală ca noțiune nu și-ar mai avea rostul). Deci, pentru a surprinde informația intrinsecă a unei probe, pe lângă tendința centrală, trebuie descrisă și variabilitatea acesteia.

În general, variabilitatea este surprinsă de modul în care valorile individuale ale unei variabile „gravitează" în jurul tendinței centrale sau se distribuie față de aceasta.

Cei mai comuni descriptori ai variabilității unei probe sunt **amplitudinea** și **deviația standard**.

Amplitudinea (ω). Este o măsură simplu de calculat a dispersiei datelor dintr-o probă. Amplitudinea reprezintă diferența dintre valoarea maximă (x_{max}) și valoarea minimă (x_{min}) dintr-o probă, adică diferența dintre cele două limite ale **intervalului de variație**.

$$\omega = x_{max} - x_{min} \qquad \text{intervalul de variație} = [x_{min}, x_{max}]$$

Când datele sunt sub forma claselor de distribuție a frecvențelor și valorile extreme (x_{max} și x_{min}) nu se cunosc cu exactitate, amplitudinea se calculează ca diferența dintre centrele intervalelor primei și ultimei clase. Dacă ne referim la exemplul 2.3 și presupunem că avem la dispoziție doar tabelul 2.5, atunci amplitudinea poate fi apreciată astfel:

$$\omega = 369{,}5 - 152{,}5 = 217 .$$

Este măsura cea mai utilă a variabilității, atunci când o decizie este condiționată de valorile extreme ale unei variabile. Prezintă însă și o serie de neajunsuri: depinde doar de valorile extreme care adesea sunt excepționale, prezintă fluctuații mari de la o probă la alta și nu este influențată de simetria repartiției dintre extreme (poate avea aceeași valoare pentru o repartiție

simetrică și pentru una puternic asimetrică).

Deviația standard (s, σ). Este descriptorul variabilității cel mai frecvent și mai util, care se folosește în analiza statistică a datelor. Atunci când valoarea sa se obține prin folosirea tuturor valorilor individuale ale unei variabile dintr-o populație, se numește **deviația standard a populației** și se notează cu σ. Dacă se calculează pornind de la datele dintr-o probă extrasă din populația de cercetat, se numește **deviația standard a probei** și se notează cu s. În ambele cazuri, deviația standard reprezintă media abaterilor valorilor individuale față de valoarea mediei. Abaterea unei valori individuale față de medie poate fi scrisă ca diferența dintre valoarea respectivă și valoarea mediei $(x - \bar{x})$. În continuare, pentru a afla media abaterilor, ar trebui ca abaterile tuturor valorilor să se însumeze și apoi să se împartă la numărul valorilor individuale luate în analiză (n sau N). Totuși există o problemă cu privire la acest raționament:

$$\sum(\bar{x} - x) = \sum\left(\frac{\sum x}{n}\right) - \sum x = n \cdot \left(\frac{\sum x}{n}\right) - \sum x = \sum x - \sum x = 0\,.$$

Deci suma abaterilor este zero, ceea ce reprezintă un impas în calcularea mediei abaterilor, aceasta fiind egală cu $0/n = 0$. Aceasta ar însemna că valorile individuale se confundă cu cea a medie, ceea ce este în majoritatea cazurilor (dacă nu în toate) imposibil, din cauza variabilității naturale. Ca urmare, pentru a depăși acest impas, este nevoie să pozitivăm toate abaterile printr-un procedeu reversibil care să nu influențeze rezultatul final. Un astfel de procedeu constă în ridicarea abaterilor la pătrat. Se obține astfel **suma pătratelor abaterilor** care, împărțită la numărul de valori, dă media aritmetică a pătratelor abaterilor sau **varianța** (s^2, σ^2).

$$s^2 = \frac{\sum(x-\bar{x})^2}{n} \qquad \text{sau} \qquad \sigma^2 = \frac{\sum(x-\bar{x})^2}{N}$$

În cazul variației, unitățile de apreciere ale variabilei se modifică – de exemplu, mm devin mm^2, g devine g^2 –, pierzându-și sensul. Pentru a reveni la unitățile originale de măsură, se extrage radicalul din valoarea variației și se obține astfel deviația standard.

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} \qquad \text{sau} \qquad \sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{N}}$$

Printr-o rearanjare algebrică a formulelor de mai sus se poate obține una mai ușor de utilizat în practică.

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}} \qquad \text{sau} \qquad \sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$$

Atunci când deviația standard a probei se utilizează pentru estimarea deviației standard a populației (se estimează σ pe baza s), suma pătratelor abaterilor se împarte la $n-1$ și nu la n, ceea ce are drept efect creșterea valorii s.

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \quad \text{sau formula cea mai uzuală} \quad s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

Creșterea valorii s reprezintă o marjă ce trebuie luată pentru folosirea mediei probei $\bar{x}$ (o statistică) în loc de media populației μ (un parametru) pentru estimarea deviației standard populaționale σ. Valoarea $\boldsymbol{n-1}$ se numește în statistică **numărul gradelor de libertate** și reprezintă un concept general desemnând numărul de elemente independente (variabile aleatoare, statistici etc.) pentru a defini starea unui sistem sau a unui ansamblu (numărul de elemente considerate simultan minus numărul de relații independente care le leagă). De exemplu, dacă presupunem că avem o probă $n = 10$, cu $\bar{x} = 20$, și trebuie să „inventăm" valorile observațiilor, avem libertatea să desemnăm primele 9 valori (adică $n-1 = 10-1 = 9$). A zecea valoare nu mai avem libertatea să o inventăm, deoarece aceasta trebuie să fie o valoare care să permită obținerea mediei egale cu 20. Cu alte cuvinte, media constituie o relație independentă, o constrângere a valorilor din probă. Deci, dacă media este 20, suma celor 10 valori va fi $20 \cdot 10 = 200$. Desemnăm primele patru valori: 26, 18, 14, 25, 20, 28, 35, 17 și 7. Suma lor este 190. Pentru ca suma să fie 200, înseamnă că al zecelea număr

poate fi doar 10 (200– 190 = 10) și astfel media să fie 20 (200/10 = 20).

Deviațiile standard ale probelor pot fi utilizate pentru a compara variabilitatea acestora. Compararea directă a deviațiilor standard obținute în urma analizei unor probe cu medii diferite (de exemplu, o medie de ordinul zecilor și o alta de ordinul sutelor) nu are nici o valoare. Astfel, pentru compararea variabilității probelor din populații cu medii diferite, se folosește **coeficientul de variație (*CV*)**.

$$CV = \frac{s}{\bar{x}} \qquad \text{sau} \qquad CV\% = \frac{s}{\bar{x}} \cdot 100$$

Exemplul 3.1. Într-un studiu s-a urmărit înălțimea fitoindivizilor de migdal pitic (*Amygdalus nana*). În acest sens s-a măsurat înălțimea a 50 de tulpini în cm dintr-un pâlc compact. S-a realizat apoi descrierea statistică a probei.

59,3	59,6	60,3	60,6	60,9	62,6	62,7	62,8	63,7	63,7
65,5	65,8	66,6	67,1	67,2	67,5	67,8	68,1	68,9	69
69	69	69	69,4	69,6	70	70,2	71,1	71,4	71,6
71,9	72,6	73,4	73,6	74,1	75	75,1	75,2	76,3	76,9
77	77,5	77,5	77,7	78,8	79,3	81,5	82,8	86,7	89,6

Valoarea cea mai frecvent întâlnită este 69, deci $Mo = 69$.

Mediana va fi egală cu media celor două valori centrale, deoarece $n = 50$, un număr par.

$$Me = \frac{\frac{x_{50}}{2} + \frac{x_{50+1}}{2}}{2} = \frac{69{,}6+70}{2} = 69{,}8$$

Media este suma tuturor valorilor împărțită la 50.

$$\bar{x} = \frac{3542{,}5}{50} = 70{,}85$$

Amplitudinea este dată de diferența dintre valoarea maximă și cea minimă din probă.

$\omega = 89{,}6 - 59{,}3 = 30{,}3$

Deviația standard se calculează conform formulei utilizate pentru estimarea deviației standard populaționale.

$$s = \sqrt{\frac{253329{,}99 - \frac{12549306{,}25}{50}}{50-1}} = \sqrt{\frac{253329{,}99 - 250986{,}13}{50-1}} = \sqrt{\frac{2343{,}87}{49}} =$$

$$= \sqrt{47{,}83} = 6{,}92$$

$$s^2 = 47{,}83$$

Coeficientul de variație se află împărțind deviația standard la medie.

$CV = \frac{6{,}92}{70{,}85} = 0{,}098$, adică 9,8% .

Datele din proba cercetată prezintă o distribuție simetrică, pentru că cele trei măsuri ale tendinței centrale au valori foarte apropiate.

4. DISTRIBUȚII PROBABILISTICE

Teoria matematică a probabilităților a apărut în urma studiului jocurilor și arată care va fi rezultatul, în general, dacă se extrag probe în mod repetat din aceeași populație statistică.

Probabilitatea de apariție a unui anumit eveniment reprezintă șansa ca evenimentul respectiv să se întâmple, exprimată de la 0 la 1 sau de la 0 la 100%. O probabilitate apropiată de 1 înseamnă că evenimentul este unul probabil, iar o probabilitate apropiată de 0 înseamnă că evenimentul respectiv este puțin probabil.

Există mai multe modalități de aflare a probabilității unui eveniment, dintre care două sunt mai des utilizate. Prima modalitate este cea empirică, bazată pe cunoștințe anterioare cu privire la evenimentul respectiv în populație. De exemplu, dacă se știe că într-o populație doi din 5 indivizi aparțin sexului masculin, atunci se poate spune că probabilitatea ca un individ selectat la întâmplare din populație să fie mascul este de 2/3 sau 0,67 (sau 67%). Pentru a afla această probabilitate, sunt necesare cunoștințe asupra raportului dintre sexe în populația studiată. A doua modalitate de aflare a probabilității este cea bazată pe considerații teoretice. De exemplu, probabilitatea de a obține un anumit număr prin aruncarea unui zar este de 1/6 adică 0,1667. În acest caz, nu este nevoie ca zarul să fie aruncat pentru a ajunge la acest rezultat.

În cele două situații de mai sus, probabilitatea a fost calculată sub forma unui raport. Acest aspect este surprins de **regula împărțirii** conform căreia probabilitatea unui eveniment este dată de numărul de posibilități în care evenimentul respectiv poate să apară împărțit la numărul total de evenimente ce pot să apară.

În primul exemplu de mai sus, erau două posibilități de apariție a unui mascul din trei indivizi pentru care nu se specifică sexul. În al doilea exemplu, era o posibilitate de a obține un anumit număr din șase numere posibile.

În general, se pot face operații cu probabilități, dintre care cele mai

frecvent utilizate sunt adunarea sau înmulțirea. Deoarece probabilitățile sunt fracții, adunarea probabilităților duce la o creștere a probabilității compuse, în timp ce înmulțirea, la o scădere a acesteia.

Dacă se cunoaște probabilitatea de apariție a unui rezultat A și cea de apariție a unui rezultat B, probabilitatea apariției simultane a ambelor rezultate este, conform **regulii înmulțirii,** egală cu produsul probabilităților individuale: $p(A \text{ și } B) = p(A) \cdot p(B)$. De exemplu, dacă probabilitatea ca dintr-un ou să iasă o anumită specie este de 0,2 și probabilitatea ca respectivul individ să fie mascul este de 0,5, atunci probabilitatea ca din ou să apară un individ mascul din specia de interes va fi egală cu produsul probabilităților individuale ale celor două rezultate: $0{,}2 \cdot 0{,}5 = 0{,}10$.

Dacă se cunoaște probabilitatea de apariție a unui rezultat A și cea de apariție a unui rezultat B, probabilitatea apariției unuia din cele două rezultate la un moment dat este, conform **regulii adunării,** egală cu suma probabilităților individuale: $p(A \text{ sau } B) = p(A) + p(B)$. De exemplu, dacă probabilitatea ca dintr-un ou să iasă o anumită specie este de 0,2 și probabilitatea de apariție a unei alte specii este 0,3, atunci probabilitatea ca din ou să iasă prima sau a doua specie va fi egală cu suma probabilităților individuale ale celor două rezultate: $0{,}2 + 0{,}3 = 0{,}5$.

O **distribuție probabilistică** este o distribuție a probabilităților similară cu o distribuție a frecvențelor (secțiunea 2.2), cu deosebirea că prima redă probabilitatea de apariție a evenimentelor și nu frecvența acestuia. Deci distribuțiile probabilistice se bazează pe probabilitățile, calculate pornind de la anumite premise ca evenimentele să apară și nu pe frecvențele observate ale evenimentelor. O distribuție a frecvențelor poate fi convertită la o distribuție a probabilităților, dacă fiecare frecvență este convertită la probabilitate prin împărțirea la numărul total de observații (dimensiunea probei).

Utilitatea distribuțiilor probabilistice este multiplă: permite estimarea probabilității ca un anumit eveniment să aibă un anumit rezultat și poate fi folosită pentru a calcula o distribuție de frecvențe de așteptat (estimate). Așa cum o probabilitate poate fi estimată prin împărțirea frecvenței unei anumite observații la numărul total de observații, tot așa, o frecvență estimată poate fi calculată prin înmulțirea probabilității estimate cu numărul total de observații.

Astfel, se pot compara frecvențele observate cu cele estimate, după

un anumit model. Dacă diferențele dintre cele două tipuri de frecvențe nu sunt semnificative, atunci modelul după care s-au calculat frecvențele estimate este valabil și pentru cele observate.

Modelele folosite în calcularea probabilităților teoretice sunt modele matematice. Dintre acestea, o parte au o importanță practică deosebită pentru cercetarea ecologică. Astfel, există trei distribuții probabilistice asociate cu variabile discrete (caractere numărabile), folosite drept model în studiile ecologice: **distribuția binomială, distribuția Poisson** și **distribuția binomială negativă**. Dintre distribuțiile probabilistice folosite pentru variabile continue (caractere măsurabile), probabil cea mai importantă, mai ales din punct de vedere conceptual, este **distribuția normală**.

4.1. DISTRIBUȚIA BINOMIALĂ

Această distribuție are următoarele particularități:

1. Observațiile sunt sub formă de număr de entități;
2. O observație se poate clasifica în una din două categorii posibile, distincte (mascul sau femelă, specia A sau ≠A, adult sau ≠adult);
3. Varianța unei probe de frecvențe este mai mică decât media;
4. Dispersia entităților numărate este uniformă.

Dacă se ia în considerare sexul unui individ dintr-o populație, această variabilă poate lua doar două valori: mascul sau femel. Dacă se selectează aleator un singur individ dintr-o populație, șansele ca acesta să fie mascul sunt egale cu șansele să fie femelă. Cum rezultatul poate fi doar unul singur, înseamnă că probabilitatea de a obține un mascul este $1/2 = 0{,}05$, iar probabilitatea de a obține o femelă este tot $1/2 = 0{,}05$. Dacă p este probabilitatea de a obține mascul și q este probabilitatea de a obține femelă, atunci suma probabilităților va fi $p + q = 1$, iar $p = 1 - q$ și $q = 1 - p$. Distribuția probabilistică în cazul sexului unui individ este reprezentată în figura 4.1.

Dacă se extrag doi indivizi din populație, atunci există patru rezultate posibile (tab. 4.1). Probabilitățile fiecărui rezultat pot fi obținute conform regulii înmulțirii. Astfel, probabilitatea de a obține doi masculi va

fi $0,5 \cdot 0,5 = 0,5^2 = 0,25$. Probabilitatea de a obține două femele se calculează în același fel. Probabilitatea de a obține un mascul și o femelă este egală cu probabilitatea de a obține o femelă și apoi un mascul, adică $0,5 \cdot 0,5 = 0,25$.

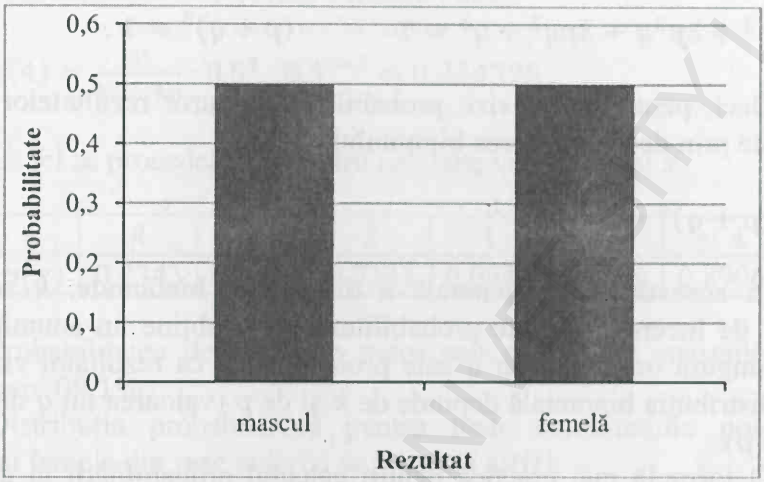


Figura 4.1. Distribuția probabilistică a sexului în cazul unui individ

Tabelul 4.1. Probabilitățile diferitelor rezultate la extragerea a doi indivizi

Individul 1 / Individul 2	Mascul (M) $p = 0,5$	Femelă (F) $q = 0,5$
Mascul (M) $p = 0,5$	MM $p \cdot p = 0,25$	FM $q \cdot p = 0,25$
Femelă (F) $q = 0,5$	MF $p \cdot q = 0,25$	FF $q \cdot q = 0,25$

Dacă nu se ia în considerație ordinea în cadrul rezultatului mascul-femelă, probabilitatea obținerii unui mascul și a unei femele va fi dată, conform regulii adunării, de suma dintre probabilitatea de a obține mascul-femelă și probabilitatea de a obține femelă-mascul, adică $pq + qp = 0,25 + 0,25 = 0,5$.

Suma probabilităților tuturor rezultatelor este 1. Dacă se generalizează suma probabilităților tuturor rezultatelor, atunci se obține:

$p^2 + 2pq + q^2 = 1$ $(p+q)^2 = 1$.

Dacă se repetă distribuția probabilistică pentru trei indivizi, atunci relația generală devine:

$p^3 + 3p^2q + 3pq^2 + q^3 = 1$ $(p+q)^3 = 1$.

Deci, pentru k indivizi, probabilitățile tuturor rezultatelor posibile vor fi date prin descompunerea binomului:

$(p+q)^k = 1$.

În această relație generală a distribuției binomiale, k reprezintă numărul de încercări, p este probabilitatea de a obține un anumit rezultat dintr-o singură încercare, iar q este probabilitatea ca rezultatul vizat să nu apară. Distribuția binomială depinde de k și de p (valoarea lui q depinde de cea a lui p).

O formulă mai practică pentru calculul probabilității în distribuția binomială este:

$p(x) = \frac{k!}{x!(k-x)!} \cdot p^x \cdot q^{k-x}$,

$p(x)$ – probabilitatea de a obține un anumit număr de rezultate;
x – numărul de rezultate de interes;
k – numărul de evenimente sau încercări;
p – probabilitatea de a obține un rezultat de interes;
q – probabilitatea de a nu obține un rezultat de interes.

Exemplul 4.1. Dacă o pontă conține 6 ouă, iar probabilitatea ca dintr-un ou să apară un mascul este 0,5, care sunt probabilitățile de a obține 4 sau mai puțini masculi?

În această problemă datele sunt următoarele:
$k = 6$ $p = 0{,}5$
$q = 1 - p = 1 - 0{,}5 = 0{,}5$
$x \leq 4$.

Aceasta înseamnă că probabilitatea de a obține 4 sau 3 sau 2 sau 1 sau 0 masculi va fi, conform regulii adunării, suma probabilităților $p(4) + p(3) + p(2) + p(1) + p(0)$.

Probabilitatea de a obține 4 masculi este:

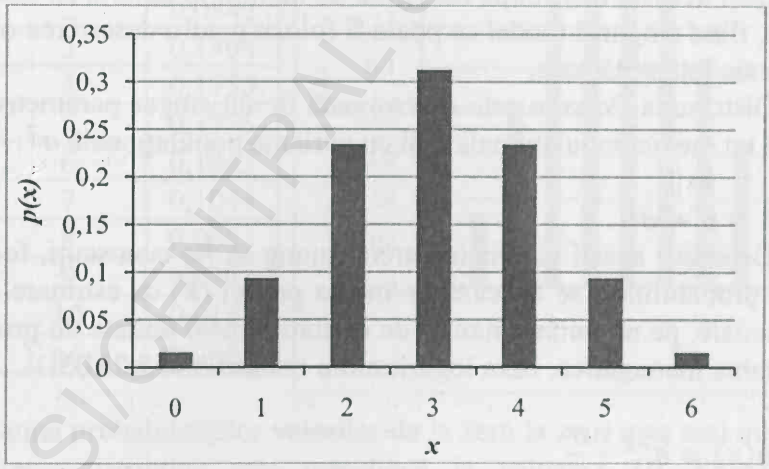
$$p(4) = \frac{6!}{4!(6-4)!} \cdot 0{,}5^4 \cdot 0{,}5^{6-4} = 0{,}234375 .$$

La fel se procedează și pentru celelalte valori ale lui x.

x	4	3	2	1	0	≤ 4
$p(x)$	0,2343	0,3125	0,2343	0,0937	0,0156	0,8906

Probabilitatea de a obține patru sau mai puțini masculi este de aproximativ 89,1%.

Distribuția probabilistică pentru toate combinațiile posibile de masculi și femele din șase indivizi se prezintă astfel:



Distribuția este simetrică deoarece $p = q$. În general, cu cât k este mai mare și p are o valoare apropiată de cea a lui q, cu atât distribuția binomială este mai simetrică, iar diferențele dintre probabilități (coloanele diagramei) sunt mai mici.

4.2. DISTRIBUȚIA POISSON

Această distribuție are următoarele particularități:

1. Observațiile sunt sub formă de număr de entități;
2. Observațiile se obțin din unități de probă definite (pătrate de probă, intervale de timp etc.) și pot fi organizate într-o distribuție a frecvențelor;
3. Varianța probei este aproximativ egală cu media acesteia;
4. Entitățile numărate sunt relativ rare (mult mai puține decât ar putea să conțină unitatea de probă).
5. Dispersia entităților în timp și spațiu este aleatoare, ceea ce înseamnă că entitățile nici nu se atrag, nici nu se resping, adică sunt independente unele față de altele.

Deși această distribuție este greu de întâlnit în natură, ea este utilă în ecologie, fiind singurul model ce poate fi folosit pentru descrierea obiectelor cu dispersie întâmplătoare.

Distribuția Poisson este determinată de un singur parametru λ, care este egal cu media populațională μ și cu varianța populațională σ^2:

$$\lambda = \mu = \sigma^2 .$$

Deoarece acești parametri rareori ajung să fie cunoscuți, formula de calcul a probabilității se bazează pe media probei ($\bar{x}$) ca estimare a mediei populaționale, pe un anumit număr de entități dintr-o unitate de probă (x) și pe constanta matematică, baza logaritmului natural $e = 2{,}7183$:

$$p(x) = e^{-\bar{x}} \cdot \frac{\bar{x}^x}{x!} .$$

Exemplul 4.2. Să se estimeze distribuția probabilistică Poisson pentru 10 sau mai puțini indivizi pe unitate de probă, știind că numărul mediu de indivizi pe unitate de probă este 5.

Datele acestei probleme sunt:

$\bar{x} = 5$

$e = 2{,}7183$

$x \leq 10$.

Trebuie calculată probabilitatea Poisson de a identifica 10 sau mai puțini indivizi pe unitate de probă.

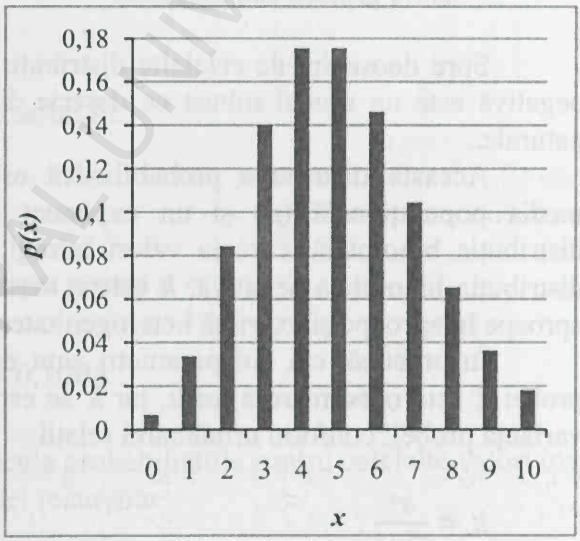
Probabilitatea de a identifica 10 indivizi pe unitate de probă este:

$$p(10) = 2{,}7183^{-5} \cdot \frac{5^{10}}{10!} = 0{,}0181 ,$$

$$p(9) = 2{,}7183^{-5} \cdot \frac{5^{9}}{9!} = 0{,}0363 .$$

La fel se procedează și pentru celelalte valori.

x	$p(x)$
0	0,0067
1	0,0337
2	0,0842
3	0,1404
4	0,1755
5	0,1755
6	0,1462
7	0,1044
8	0,0653
9	0,0363
10	0,0181
$\sum p(x)$	0,9863



Suma probabilităților valorilor de la zero la zece este mai mică decât 1, deoarece distribuția este trunchiată la valoarea 10. Probabilitățile valorilor mai mari de 10 sunt extrem de mici. Distribuția este ușor asimetrică. Dacă λ depășește valoarea 10, distribuția tinde să devină aproximativ simetrică.

4.3. DISTRIBUȚIA BINOMIALĂ NEGATIVĂ

Această distribuție are următoarele particularități:

1. Observațiile sunt sub formă de număr de entități;
2. Observațiile se obțin din unități de probă definite (pătrate de probă, intervale de timp etc.) și pot fi organizate într-o distribuție a frecvențelor;
3. Varianța probei este evident mai mare decât media acesteia;
4. Entitățile numărate sunt relativ rare (mult mai puține decât ar putea să conțină unitatea de probă).
5. Dispersia entităților nu este nici uniformă, nici aleatoare, ci poate să fie grupată sau agregată.

Spre deosebire de celelalte distribuții discrete, distribuția binomială negativă este un model robust ce descrie dispersia a numeroase populații naturale.

Această distribuție probabilistică este definită de doi parametri: media populațională (μ) și un exponent k. Spre deosebire de k din distribuția binomială, care ia valori întregi, fiind o variabilă discretă, în distribuția binomială negativă, k este o variabilă continuă, ce ia valori din aproape în aproape și exprimă heterogenitatea unei distribuții.

În practică, cei doi parametri sunt estimați pe baza probei. Media probei $\bar{x}$ este o estimare a lui μ, iar k se estimează pornind de la media și varianța probei, conform următoarei relații:

$$k = \frac{\bar{x}^2}{s^2 - \bar{x}}.$$

Probabilitățile individuale, adică probabilitățile obținerii unui anumit număr de entități x dintr-o unitate de probă, se pot afla prin descompunerea expresiei $(q - p)^{-k}$, unde $p = \bar{x}/k$, iar $q = 1 + p$.

Din motive practice, probabilitatea în distribuția binomială negativă se calculează urmând următoarele etape:

1. Se calculează valoarea k.
2. Se calculează probabilitatea pentru $x = 0$:

$$p(0) = q^{-k} = \left(1 + \frac{\bar{x}}{k}\right)^{-k}.$$

3. Pentru oricare $x > 0$, se calculează probabilitatea astfel:

$$p(x > 0) = \frac{k+x-1}{x} \cdot \frac{\bar{x}}{\bar{x}+k} \cdot p(x-1).$$

Exemplul 4.3. Care sunt probabilitățile ca pe niște suprafețe de probă să apară zece sau mai puțini indivizi? Media numărului de indivizi pe suprafață de probă este trei, iar deviația standard este cinci.

$\bar{x} = 3$
$s^2 = 5$

Trebuie calculată valoarea lui k:

$$k = \frac{3^2}{5-3} = 4{,}5.$$

Se calculează probabilitatea pentru $x = 0$:

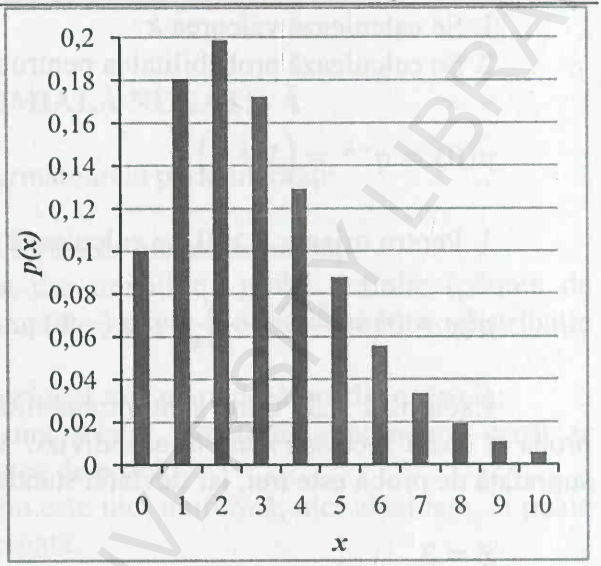
$$p(0) = \left(1 + \frac{3}{4{,}5}\right)^{-4{,}5} = 0{,}1004.$$

În continuare se pot calcula probabilitățile pentru celelalte valori mai mari ca unu, cu ajutorul formulei recurente:

$$p(1) = \frac{4{,}5+1-1}{1} \cdot \frac{3}{3+4{,}5} \cdot 0{,}1004 = 0{,}1807$$
$$p(2) = \frac{4{,}5+2-1}{2} \cdot \frac{3}{3+4{,}5} \cdot 0{,}1807 = 0{,}1988.$$

În același mod se continuă și pentru celelalte valori. Rezultatele probabilităților pentru valorile mai mici sau egale cu zece sunt trecute în tabelul următor:

x	$p(x)$
0	0,1004
1	0,1807
2	0,1988
3	0,1723
4	0,1292
5	0,0879
6	0,0556
7	0,0334
8	0,0192
9	0,0107
10	0,0058
$\sum p(x)$	0,9938



Se poate observa că suma probabilităților valorilor este mai mică decât 1. Probabilitățile pentru valorile mai mari de 10 vor fi extrem de mici. Pe măsură ce valoarea k crește și deviația standard (s^2) scade în raport cu media ($\bar{x}$), gradul de simetrie al distribuției crește. Pentru valori ale lui $k > 20$, distribuția probabilistică este aproape simetrică.

4.4. ESTIMAREA DISPERSIEI UNEI POPULAȚII

4.4.1. Indici de dispersie

Dispersia unei populații se referă la modul de repartizare a indivizilor în spațiu. Dispersia se apreciază pe baza poziției unor indivizi, relativ la poziția celorlalți.

În general, dispersia unei populații poate fi de trei tipuri: **uniformă**, **aleatoare** și **grupată** (fig. 4.2). În cele mai multe cazuri, dispersia este privită din perspectivă spațială, dar se poate investiga și dispersia în timp a unor evenimente. În primul caz, unitatea de probă poate fi un pătrat, în timp ce în al doilea caz, poate fi un interval de timp. Dacă se urmăresc organisme parazite, atunci unitatea de probă poate fi organismul gazdă.

În principiu, aprecierea dispersiei se face în funcție de variabilitatea

densității entităților pe unitate de probă. În cazul unei dispersii uniforme, densitățile entităților pe unitate de probă vor fi cam aceleași și variabilitatea acestor densități va fi relativ mică. În cazul unei dispersii grupate, densitățile vor avea valori extreme, fie foarte mari, fie aproape nule, iar variabilitatea acestor densități va fi relativ mare. În cazul dispersiei aleatoare, vor exista densități cu valori mari, mici și intermediare și, ca urmare, variabilitatea acestor densități va fi intermediară față de variabilitatea densităților din celelalte tipuri de dispersie.

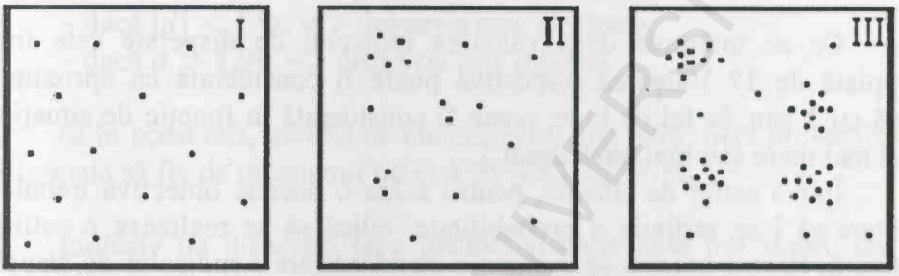


Figura 4.2. Tipuri de dispersie: I uniformă, II aleatoare, III grupată

Pornind de la acest principiu, dacă se folosește varianța densităților pe unitate de probă ca descriptor al variabilității și media densităților pe unitate de probă ca termen de comparație a magnitudinii variației, se poate calcula un **indice de dispersie** astfel:

$$\text{Indice de dispersie} = \frac{s^2}{\bar{x}}.$$

Când datele sunt prezentate sub formă de tabel de distribuție a frecvențelor, atunci formulele varianței și mediei pot fi adaptate pentru facilitarea calculelor astfel:

x	f	$x \cdot f$
x_1	f_1	$x_1 \cdot f_1$
x_2	f_2	$x_2 \cdot f_2$
...	...	...
x_k	f_k	$x_k \cdot f_k$
	$\sum f = n$	$\sum(x \cdot f)$

$$\bar{x} = \frac{\sum(xf)}{n}$$

$$s^2 = \frac{\sum[f(x-\bar{x})^2]}{n-1}.$$

Indicele de dispersie poate fi subunitar, egal cu unitatea sau supraunitar. Astfel, în funcție de valoarea raportului, se poate decide care este tipul de dispersie:

– dacă $\frac{s^2}{\bar{x}} < 1$, dispersia este uniformă;
– dacă $\frac{s^2}{\bar{x}} \approx 1$, dispersia este aleatoare;
– dacă $\frac{s^2}{\bar{x}} > 1$, dispersia este grupată.

Ce se întâmplă dacă valoarea indicelui de dispersie este foarte apropiată de 1? Valoarea respectivă poate fi considerată ca aproximativ egală cu 1 sau, la fel de bine, poate fi considerată în funcție de situație ca fiind mai mare sau mai mică decât 1.

Într-o astfel de situație, pentru a lua o decizie obiectivă trebuie ca acesteia să i se atribuie o probabilitate, adică să se realizeze o estimare statistică. Pentru aceasta se impune o standardizare a indicelui de dispersie prin înmulțirea acestuia cu numărul gradelor de libertate, adică cu numărul unităților de probă minus unu $(n-1)$, obținându-se o statistică de tip χ^2:

$$\chi^2 = (n-1) \cdot \frac{s^2}{\bar{x}} .$$

Când numărul unităților de probă este relativ mic $(n < 30)$, se compară statistica χ^2 cu valorile critice ale distribuției χ^2 pentru $n-1$ grade de libertate și pentru pragul de confidență 0,975 și, respectiv, 0,025:

– dacă $\chi^2 < \chi^2_{(0,975,n-1)}$ => dispersia este uniformă;
– dacă $\chi^2_{(0,975,n-1)} < \chi^2 < \chi^2_{(0,025,n-1)}$ => dispersia este aleatoare;
– dacă $\chi^2 > \chi^2_{(0,025,n-1)}$ => dispersia este grupată.

În oricare dintre situațiile de mai sus, nivelul de confidență al estimării este de 0,05 (cele două valori critice exclud fiecare câte 0,025 din fiecare coadă a distribuției χ^2). Deci probabilitatea ca dispersia să fie de un anumit tip este de 0,95 sau 95%.

Când numărul unităților de probă este relativ mare ($n \geq 30$), atunci se calculează o statistică d conform relației:

$$d = \sqrt{2 \cdot \chi^2} - \sqrt{2 \cdot (n-1) - 1}\,.$$

Această statistică se compară cu valoarea critică 1,96, care este o valoare z ce exclude 0,05 din distribuția normală standard (secțiunea 4.5):

– dacă $d < -1{,}96 \Rightarrow$ dispersia este uniformă;
– dacă $|d| < 1{,}96 \Rightarrow$ dispersia este aleatoare;
– dacă $d > 1{,}96 \Rightarrow$ dispersia este grupată.

Și în acest caz, nivelul de încredere este de 0,05, deci probabilitatea ca dispersia să fie de un anumit tip este de 0,95 sau 95%.

Indicele de dispersie face distincția dintre cele trei tipuri, dar nu poate indica gradul de aglomerare în cazul unei dispersii grupate, deoarece este puternic influențat de numărul total de entități individuale din toate unitățile de probă. Pentru a aprecia gradul de aglomerare dintr-o dispersie grupată, se folosește **indicele Green (*IG*)**, care elimină dependența de numărul tuturor indivizilor din toate unitățile de probă. Numărul total de indivizi din toate unitățile de probă va fi egal cu $\sum(xf)$ în cazul în care datele sunt sub formă de tabel de frecvențe sau cu $\sum(x)$ dacă datele sunt sub formă de densități pe unitatea de probă.

$$IG = \frac{\frac{s^2}{\bar{x}} - 1}{\sum(xf) - 1} \quad \text{sau} \quad IG = \frac{\frac{s^2}{\bar{x}} - 1}{\sum(x) - 1}$$

Acest indice ia valori între 0, pentru dispersie aleatoare, și 1, pentru dispersie grupată, cu grad maxim de aglomerare (toate entitățile au fost identificate într-o singură unitate de probă). Acest indice poate fi folosit pentru compararea gradului de aglomerare a unor probe diferite ca număr de entități, medie sau număr de unități de probă.

Exemplul 4.4. Într-o populație de plante, prin investigarea a 14 pătrate de probă s-au obținut densitățile indivizilor. Care este tipul de dispersie al populației?

0	0	0	5	3	48	1
2	30	5	9	22	1	0

$$\bar{x} = \frac{126}{14} = 9$$
$$s^2 = 207{,}69$$
$$\frac{s^2}{\bar{x}} = \frac{207{,}69}{9} = 23{,}08\,.$$

Indicele de dispersie este evident mai mare ca 1. În continuare vom verifica dacă dispersia aglomerată este semnificativă.

$$\chi^2 = (14-1)\cdot 23{,}08 = 300{,}04$$

Această valoare se compară cu valorile critice calculate pentru 13 grade de libertate și pentru nivelurile de semnificație 0,975 și, respectiv, 0,025 (anexa 3):

$$\chi^2_{(0{,}975,13)} = 5{,}009$$
$$\chi^2_{(0{,}025,13)} = 24{,}736\,.$$

Valoarea calculată ($\chi^2 = 300{,}04$) este mai mare decât limita superioară ($\chi^2_{(0{,}025,13)} = 24{,}736$) a intervalului de confidență pentru dispersia aleatoare, deci populația investigată are o dispersie aglomerată, cu o probabilitate de 0,95 sau 95%.

Se estimează gradul de aglomerare cu indicele Green (IG). Datele sunt sub formă de densitate pe pătrat de probă, deci numărul total de indivizi identificați în cele 14 pătrate este dat de suma numărului de indivizi din fiecare pătrat.

$$IG = \frac{23{,}08-1}{126-1} = 0{,}177$$

Având în vedere că IG ia valori între 0 și 1, se poate concluziona că gradul de aglomerare nu este destul de scăzut.

În situația în care numărul unităților de probă este mai mare de 30, se poate verifica dacă frecvențele observate ale variabilei concordă cu frecvențele estimate cu ajutorul uneia dintre distribuțiile probabilistice discrete (binomială, Poisson, binomială negativă) ce servește drept model. Practic, se calculează probabilitatea de a obține o anumită valoare a variabilei pe unitate de probă, după care se convertește în frecvență prin înmulțirea cu numărul unităților de probă. Concordanța frecvențelor observate cu cele estimate (cu modelul) se poate face prin reprezentarea grafică a ambelor frecvențe și prin testul χ^2 de concordanță (secțiunea 10.1). În secțiunile următoare ne vom referi la compararea grafică a frecvențelor.

4.4.2. Modelul binomial

Calcularea probabilității binomiale de obținere a unei valori a variabilei pe unitate de probă se bazează pe procedura descrisă în exemplul 4.1, cu deosebirea că numărul de încercări (k) se estimează ca fiind valoarea rotunjită obținută prin calcularea expresiei:

$$k = \frac{\bar{x}^2}{\bar{x} - s^2}.$$

Parametrii p și q reprezintă probabilitatea ca o anumită poziție dintr-o unitate de probă să fie ocupată de o entitate.

$$p = \frac{\bar{x}}{k}, \qquad \text{iar} \qquad q = 1 - p.$$

Pe baza acestor parametrii se calculează probabilitatea binomială ($p(x)$) de a obține o anumită valoare a variabilei (x) pe unitate de probă. Ulterior se află frecvența estimată a valorii x prin înmulțirea probabilității acesteia cu numărul unităților de probă $f' = p(x) \cdot n$.

Exemplul 4.5. S-au determinat densitățile indivizilor unei specii de plante din 50 de pătrate de probă cu o anumită suprafață. Care este tipul de dispersie al populației de plante?

Nr. indivizi pe pătrat	16	17	18	19	20	21	22
Frecvența	3	5	8	14	11	6	3

Variabila este numărul indivizilor pe pătrat de probă și o vom nota cu x. Frecvențele observate le vom nota cu f.

Pentru a calcula numărul mediu de indivizi pe pătrat ($\bar{x}$), trebuie aflat numărul total de indivizi din toate pătratele. Deoarece datele sunt prezentate sub formă de tabel de frecvență, numărul total de indivizi identificați va fi $\sum(xf)$.

x	16	17	18	19	20	21	22	Suma
f	3	5	8	14	11	6	3	50
xf	48	85	144	266	220	126	66	955

$$\bar{x} = \frac{955}{50} = 19{,}1$$

$$s^2 = \frac{3(16-19{,}1)^2}{50-1} + \frac{5(17-19{,}1)^2}{50-1} + \cdots + \frac{3(22-19{,}1)^2}{50-1} = 2{,}38$$

Indicele de dispersie este evident subunitar, deci se poate ca populația să aibă o dispersie uniformă.

$$\text{Indicele de dispersie} = \frac{2{,}38}{19{,}1} = 0{,}124$$

$$\chi^2 = 0{,}124 \cdot (50-1) = 6{,}099$$

$$d = \sqrt{2 \cdot 6{,}099} - \sqrt{2(50-1)-1} = -6{,}356$$

$-6{,}356 < -1{,}96 \Rightarrow$ dispersie uniformă semnificativă

În continuare se calculează valorile k și p:

$$k = \frac{19{,}1^2}{19{,}1-2{,}38} = 21{,}82 \approx 22$$

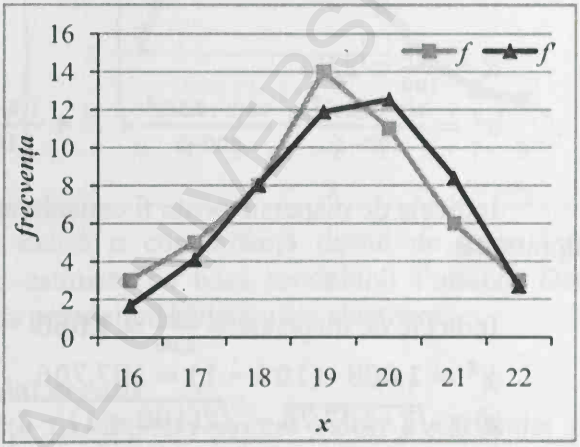
$$p = \frac{19{,}1}{22} = 0{,}876 \qquad q = 1 - 0{,}876 = 0{,}124\,.$$

Pornind de la aceste valori, se pot calcula probabilitățile pentru x indivizi la pătrat de probă (anexa 3):

$$p(16) = \frac{22!}{16!(22-16)!} \cdot 0{,}876^{16} \cdot 0{,}124^{(22-16)} = 0{,}0331 .$$

La fel se procedează și pentru celelalte valori, după care se înmulțesc cu numărul unităților de probă, obținându-se frecvențele estimate (f').

x	f	$p(x)$	f'
16	3	0,033	1,654
17	5	0,082	4,107
18	8	0,160	8,024
19	14	0,238	11,881
20	11	0,251	12,534
21	6	0,168	8,396
22	3	0,054	2,684



Se poate observa că există o concordanță destul de mare între frecvențele observate și cele estimate pe baza modelului binomial. Deci putem concluziona că dispersia populației este uniformă.

4.4.3. Modelul Poisson

Calcularea probabilității de obținere a unei valori a variabilei pe unitate de probă, conform distribuției Poisson, se bazează pe procedura descrisă în exemplul 4.2.

Exemplul 4.6. Într-un studiu s-a urmărit densitatea indivizilor unei specii de șarpe în 100 de pătrate de probă, într-o formațiune ierboasă. Ce tip de dispersie prezintă populația studiată?

Nr. indivizi/pătrat	0	1	2	3	4	5	6	7	8
Frecvența	7	16	25	18	16	10	5	2	1

Se calculează media și deviația standard la fel ca în exemplul 4.5.

x	0	1	2	3	4	5	6	7	8	Suma
f	7	16	25	18	16	10	5	2	1	100
xf	0	16	50	54	64	50	30	14	8	286

$$\bar{x} = \frac{286}{100} = 2{,}86$$

$$s^2 = \frac{7(0-2{,}86)^2}{100-1} + \frac{16(1-2{,}86)^2}{100-1} + \cdots + \frac{1(8-2{,}86)^2}{100-1} = 3{,}11$$

Indicele de dispersie poate fi considerat fie aproximativ egal cu 1, fie supraunitar.

$$\text{Indicele de dispersie} = \frac{3{,}11}{2{,}86} = 1{,}088$$

$$\chi^2 = 1{,}088 \cdot (100-1) = 107{,}706$$

$$d = \sqrt{2 \cdot 107{,}76} - \sqrt{2(100-1)-1} = 0{,}641$$

$0{,}641 < 1{,}96 \Rightarrow$ dispersie aleatoare semnificativă

În continuare se calculează probabilitatea Poisson ($p(x)$) pentru toate valorile variabilei x (anexa 3). Se obțin astfel probabilitățile de apariție a x indivizi la pătrat de probă. Probabilitatea de a obține x indivizi în 100 de pătrate se obține înmulțind $p(x)$ cu 100. Valorile astfel obținute reprezintă frecvențele teoretice (f').

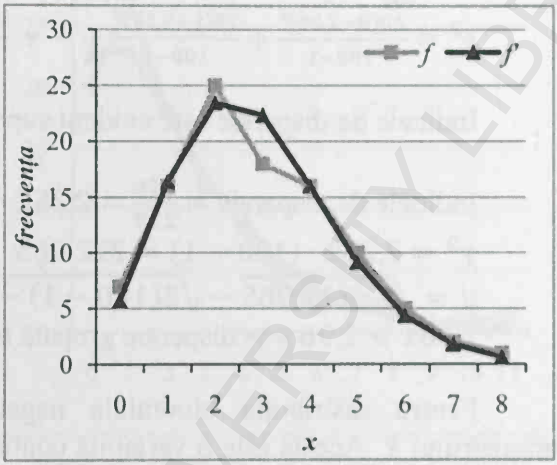
$$p(0) = 2{,}7183^{-2{,}86} \cdot \frac{2{,}86^0}{0!} = 0{,}05727$$

$$p(1) = 2{,}7183^{-2{,}86} \cdot \frac{2{,}86^1}{1!} = 0{,}16379$$

La fel se procedează și pentru celelalte valori, după care se înmulțesc cu numărul unităților de probă, obținându-se frecvențele estimate (f').

$f_1' = 0{,}05727 \cdot 100 = 5{,}727$
$f_2' = 0{,}16379 \cdot 100 = 16{,}379$.

x	f	$p(x)$	f'
0	7	0,057	5,727
1	16	0,164	16,379
2	25	0,234	23,422
3	18	0,223	22,329
4	16	0,160	15,965
5	10	0,091	9,132
6	5	0,044	4,353
7	2	0,018	1,778
8	1	0,006	0,636



Se poate observa că există o concordanță destul de mare între frecvențele observate și cele estimate pe baza modelului Poisson. Deci putem concluziona că dispersia populației studiate este aleatoare.

4.4.4. Modelul binomial negativ

Calcularea probabilității de obținere a unei valori a variabilei pe unitate de probă, conform distribuției binomiale negative, se bazează pe procedura descrisă în exemplul 4.3.

Exemplul 4.7. Într-un studiu s-a urmărit densitatea indivizilor unei specii de plante în 100 de pătrate de probă. Care este tipul de dispersie al populației?

Nr. indivizi/pătrat	0	1	2	3	4	5	6	7	8	9	10	11
Frecvența	20	27	18	12	10	4	2	3	2	0	1	1

Se calculează media și deviația standard la fel ca în exemplul 4.5.

x	0	1	2	3	4	5	6	7	8	9	10	11	Suma
f	20	27	18	12	10	4	2	3	2	0	1	1	100
xf	0	27	36	36	40	20	12	21	16	0	10	11	229

$$\bar{x} = \frac{229}{100} = 2{,}29$$
$$s^2 = \frac{20(0-2{,}29)^2}{100-1} + \frac{27(1-2{,}29)^2}{100-1} + \cdots + \frac{1(11-2{,}86)^2}{100-1} = 5{,}16$$

Indicele de dispersie este evident supraunitar.

$$\text{Indicele de dispersie} = \frac{5{,}16}{2{,}29} = 2{,}252$$
$$\chi^2 = 2{,}252 \cdot (100-1) = 222{,}965$$
$$d = \sqrt{2 \cdot 222{,}965} - \sqrt{2(100-1)-1} = 7{,}081$$
$$7{,}081 > 1{,}96 \Rightarrow \text{dispersie grupată semnificativă}$$

Pentru distribuția binomială negativă este necesară calcularea parametrului k. Acesta este o variabilă continuă, motiv pentru care valoarea sa nu se rotunjește ca în cazul distribuției binomiale.

$$k = \frac{2{,}29^2}{5{,}16-2{,}29} = 1{,}829$$

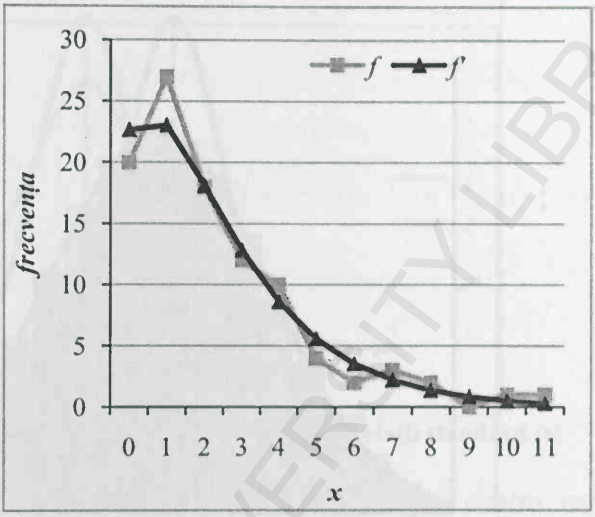
În continuare se calculează probabilitatea Poisson ($p(x)$) pentru toate valorile variabilei x:

$$p(0) = \left(1 + \frac{2{,}29}{1{,}829}\right)^{-1{,}829} = 0{,}2265\ ;$$
$$p(1) = \frac{(1{,}829+1-1)}{1} \cdot \frac{2{,}29}{2{,}29+1{,}829} \cdot 0{,}2265 = 0{,}2304\ ;$$
$$p(2) = \frac{(1{,}829+2-1)}{2} \cdot \frac{2{,}29}{2{,}29+1{,}829} \cdot 0{,}2304 = 0{,}1811\ .$$

La fel se procedează pentru restul valorilor lui x. Probabilitățile de apariție a x indivizi la pătrat de probă se înmulțesc cu numărul unităților de probă (100), obținându-se astfel frecvențele estimate (f') conform modelului binomial negativ.

$$f_1' = 0{,}227 \cdot 100 = 22{,}65$$
$$f_1' = 0{,}2304 \cdot 100 = 23{,}04$$
$$f_1' = 0{,}1815 \cdot 100 = 18{,}15$$

x	f	$p(x)$	f'
0	20	0,2265	22,65
1	27	0,2304	23,04
2	18	0,1811	18,11
3	12	0,1285	12,85
4	10	0,0863	8,63
5	4	0,0559	5,59
6	2	0,0354	3,54
7	3	0,0220	2,20
8	2	0,0135	1,35
9	0	0,0082	0,82
10	1	0,0049	0,49
11	1	0,0029	0,29



Se poate observa că există o concordanță destul de mare între frecvențele observate și cele estimate pe baza modelului distribuției binomiale negative. Deci se poate concluziona că dispersia populației studiate este aglomerată.

4.5. DISTRIBUȚIA NORMALĂ

Distribuția normală este una dintre distribuțiile continue. Aceasta descrie, mai mult sau mai puțin, distribuția unui mare număr de variabile, motiv pentru care reprezintă o bază conceptuală pentru multe procedee de analiză statistică.

Variabilele continue pot lua orice valoare între anumite limite. Dacă se reprezintă grafic distribuția frecvențelor unei astfel de variabile într-o populație, prin intermediul unei linii continue, aceasta va avea o formă simetrică, de clopot. De aceea, această distribuție mai este numită și „clopotul lui Gauss", care este unul din autorii (Moivre 1733, Legendre 1805, Laplace 1812) care a descris riguros această distribuție (Gauss 1809). Teoretic, dacă se realizează histograma folosind un număr infinit de valori individuale, iar intervalul de clasă este cel mai mic posibil, histograma sau poligonul frecvențelor tinde să devină o curbă continuă (fig. 4.3).

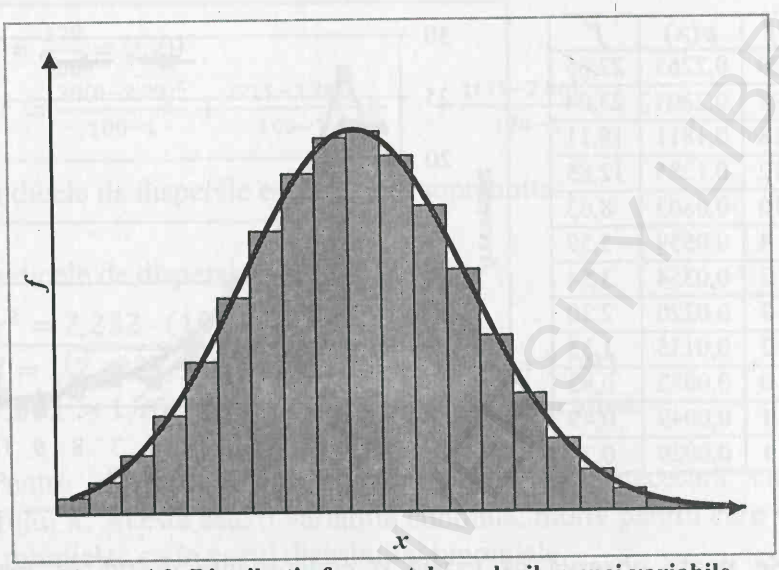


Figura 4.3. Distribuția frecvențelor valorilor unei variabile

Numeroase variabile continue întâlnite în natură au o distribuție normală. De asemenea, multe variabile care au o amplitudine mare a valorilor prezintă o distribuție aproximativ normală. Unele distribuții discrete tind să devină aproximativ normale sau simetrice pe măsură ce parametrii legați de numărul de valori cresc.

Proprietățile distribuției normale

1. Distribuția este definită de medie (μ) și de deviația standard (σ). Poziția distribuției pe abscisă este determinată de valoarea mediei, iar lărgimea acesteia, de deviația standard (fig. 4.4). Cum acești parametri pot avea o infinitate de valori diferite, înseamnă că există un număr infinit de distribuții normale.
2. Înălțimea curbei față de ordonată este dată de funcția de repartiție $f(x)$ pentru fiecare valoare individuală a variabilei:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

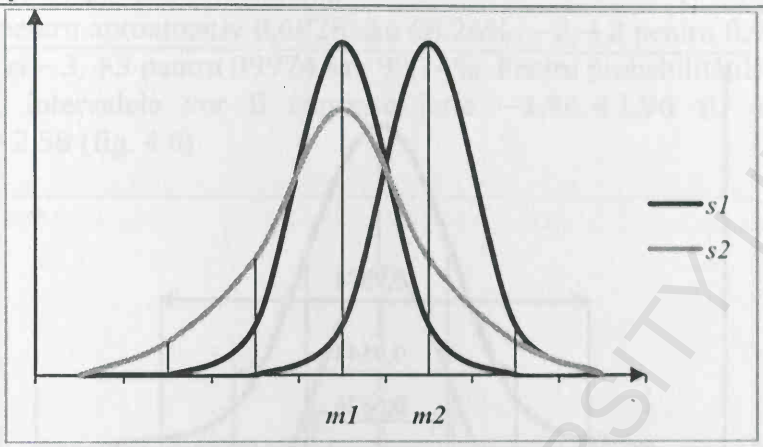


Figura 4.4. Distribuții normale diferite după medii (m) și deviații standard (s)

3. Curba este perfect simetrică față de medie, motiv pentru care media și mediana sunt egale în distribuția normală. De asemenea, valorile variabilei egale cu media sunt cele mai frecvente și astfel media este egală cu modul valorilor individuale. În concluzie, media, mediana și modul valorilor unei variabile normal distribuite sunt egale.
4. Curba distribuției cuprinde probabilitatea totală, adică 1. Dacă se consideră suprafața delimitată de curbă ca fiind 100%, atunci suprafața delimitată de valoarea $\mu - \sigma$ și $\mu + \sigma$ reprezintă aproximativ 68,26% din total. Adică în jur de 68% dintre valorile variabilei sunt cuprinse în acest interval sau probabilitatea ca o valoare selectată aleator din populație să fie cuprinsă de acest interval este de 0,68. Între $\mu - 2\sigma$ și $\mu + 2\sigma$ se găsește 95,44% din suprafața curbei, adică probabilitatea de a observa o valoare din acest interval este de 0,9544, iar intervalul $\mu - 3\sigma$, $\mu + 3\sigma$ cuprinde 99,74% din valorile individuale sau probabilitatea de a extrage o valoare din acest interval este de 0,9974 (fig. 4.5). În practică se folosesc probabilitățile 0,95 și 0,99, pentru care intervalele sunt $\mu \pm 1{,}96\sigma$ și, respectiv, $\mu \pm 2{,}58\sigma$. Probabilitatea ca o valoare să fie în afara celor două intervale va fi $1 - 0{,}95 = 0{,}05$ și, respectiv, $1 - 0{,}99 = 0{,}01$. Aceste proprietăți pot fi folosite pentru aprecierea posibilităților de apariție a unor rezultate.

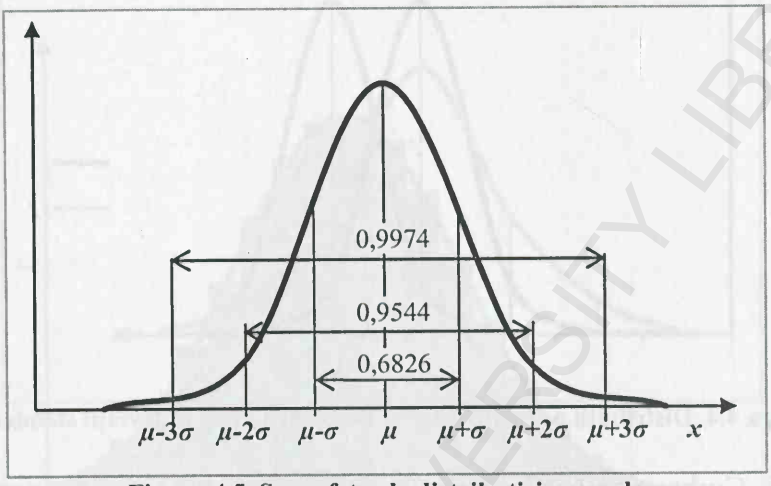


Figura 4.5. Suprafețe ale distribuției normale

Distribuția normală standard

Orice distribuție normală particulară ($N(\mu, \sigma)$) poate fi convertită la distribuția normală standard care se caracterizează prin faptul că are media zero și deviația standard unu ($N(0,1)$). Această conversie se realizează prin calcularea **scorului** z pentru fiecare valoare individuală a variabilei x, în funcție de media și de deviația standard populațională (μ și σ):

$$z = \frac{x-\mu}{\sigma}.$$

În practică, parametrii populaționali nu pot fi cunoscuți cu exactitate, caz în care pot fi substituiți cu statisticile probei ($\bar{x}$ și s), cu condiția ca dimensiunea probei sa fie mai mare sau egală cu 30 ($n \geq 30$).

$$z = \frac{x-\bar{x}}{s}$$

.

Astfel, pentru valorile x mai mici decât media, z va avea o valoare negativă, iar pentru cele mai mari decât media, o valoare pozitivă. Scorul z arată la ce distanță de medie există o anumită valoare, unitatea de referință fiind deviația standard.

Intervalele de probabilitate ale distribuției normale standard vor fi

$-1, +1$ pentru aproximativ 0,6826 sau 68,26%, $-2, +2$ pentru 0,9544 sau 95,44% și $-3, +3$ pentru 09974 sau 99,74%. Pentru probabilitățile de 0,95 și 0,99, intervalele vor fi cuprinse între $-1{,}96, +1{,}96$ și, respectiv, $-2{,}58, +2{,}58$ (fig. 4.6).

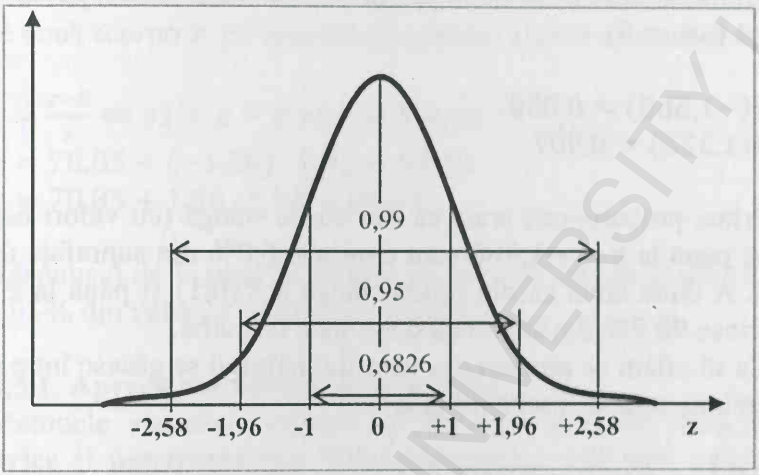


Figura 4.6. Suprafețe ale distribuției normale standard

Probabilitățile pentru diferite valori ale lui z sunt tabelate (anexa 2) sau pot fi calculate (anexa 3) pornind de la coada din stânga (cu valori extreme negative) a $N(0,1)$ și până la valoarea z calculată.

Exemplul 4.8. Considerând datele din exemplul 3.1 reprezentând înălțimile a 50 de indivizi de migdal pitic (*Amygdalus nana*), să se estimeze:

a) Ce procent din populație are înălțimea cuprinsă între 60 cm și 80 cm?

b) Între ce interval de înălțime sunt cuprinse 95% din valorile înălțimii în populație?

Se cunoaște că $\bar{x} = 70{,}85$ și $s = 6{,}92$.

Pentru a răspunde la prima întrebare trebuie făcută conversia celor două valori la distribuția normală standard, adică trebuie calculate scorurile z pentru cele două înălțimi, după care se află probabilitățile sau procentele celor corespunzătoare. Simplificând, se parcurge următoarea schemă $x \rightarrow z \rightarrow p$.

Pentru x =60 scorul $z = \frac{60-70,85}{6,92} = \frac{-10,85}{6,92} = -1,568$.
Pentru x =80 scorul $z = \frac{80-70,85}{6,92} = \frac{9,15}{6,92} = 1,322$.

Probabilitățile celor două scoruri z se caută în tabel (anexa 2) sau se calculează (anexa 3)

$$p(-1,568) = 0,058$$
$$p(1,322) = 0,907 .$$

Prima probabilitate arată că din coada stângă (cu valori negative) a $N(0,1)$ și până la $z = -1,568$ sunt cuprinse 5,8% din suprafața delimitată de curbă. A doua arată că din coada stângă a $N(0,1)$ și până la $z = 1,322$ sunt cuprinse 90,7% din suprafața delimitată de curbă.

Ca să aflăm ce procent din valorile înălțimii se găsesc între 60 cm și 70 cm, trebuie scăzute probabilitățile:

$$0,907 - 0,058 = 0,849 .$$

Răspunsul la prima întrebare este că între 60 cm și 70 cm sunt cuprinse aproximativ 85% din valorile înălțimii în populație.

Pentru a răspunde la a doua întrebare este nevoie să se parcurgă schema de raționament de la prima, dar în sens invers: $p \rightarrow z \rightarrow x$.

Scorul z pentru o anumită probabilitate se află din tabel (anexa 2) sau se calculează (anexa 3). Având în vedere că atât tabelul din anexa 2, cât și funcția din anexa 3 consideră probabilitatea din coada stângă a $N(0,1)$, înseamnă că aflarea scorului z pentru $p = 0,95$ va returna limita superioară a unui interval dispus asimetric față de medie, limita inferioară fiind către $-\infty$. Astfel, pentru a obține scorurile z care să delimiteze o suprafață de 0,95 din $N(0,1)$ simetrică față de medie, trebuie ca diferența $1 - 0,95 = 0,05$ să fie împărțită în mod egal în ambele cozi ale distribuției (ca în fig. 4.6). Deci se află scorurile z pentru $p = 0,025$ și, respectiv, pentru $p = 1 - 0,025 = 0,975$. La fel ca la punctul a), primul z separă 0,025 până în coada stângă, al doilea, 0,975 până în coada stângă, iar între cei doi se găsește o suprafață ce reprezintă 0,95 din suprafața totală de sub clopot.

$$z_{(0,025)} = -1,96$$
$$z_{(0,975)} = +1,96$$

Mai departe se calculează cele două valori ale înălțimii (x) pornind de la cele două scoruri z, pe baza relației pentru aflarea scorurilor z.

$$z = \frac{x-\bar{x}}{s} \Leftrightarrow zs = x - \bar{x} \Leftrightarrow x = \bar{x} + zs$$
$$x = 70,85 + (-1,96) \cdot 6,92 = 57,29$$
$$x = 70,85 + 1,96 \cdot 6,92 = 84,41$$

Răspunsul de la punctul b) este că între 57,29 *cm* și 84,41 *cm* sunt cuprinse 95% din valorile înălțimii în populație.

4.5.1. Aprecierea normalității datelor

Metodele statistice utilizate în ecologie sunt de două categorii: **parametrice** și **neparametrice**. Cele parametrice sunt mai puternice, dar totodată mai restrictive în sensul că se pot aplica doar dacă datele întrunesc o serie de condiții. O condiție comună pentru toate testele parametrice este ca datele să fie aproximativ normal distribuite. Metodele neparametrice nu prevăd această condiție și din această perspectivă se mai numesc și independente de distribuție. Ele pot fi utilizate pentru o gamă mai variată de situații, dar sunt mai puțin puternice (secțiunea 5.5, Erori statistice).

Deci, pentru a putea utiliza metode statistice parametrice, trebuie să se verifice normalitatea distribuției datelor. Din start trebuie subliniat că variabilele trebuie să fie apreciate pe o scală de interval sau raport, cu alte cuvinte trebuie să fie continue sau discrete (în cazul celor discrete trebuie să existe un număr relativ mare de valori posibile).

În statistică există teste dedicate, care verifică dacă datele au o distribuție aproximativ normală, dar care sunt relativ complicate și rareori utilizate, motiv pentru care doar le menționăm pe cele mai cunoscute: testul Shapiro-Wilk, testul Kolmogorov-Smirnov, testul Cramér-von-Mises, testul Jarque-Bera.

Testarea normalității se poate face și cu ajutorul testului χ^2 de concordanță între frecvențele valorilor variabilelor din probă și frecvențele estimate pe baza funcției de repartiție a distribuției normale. Acest procedeu

este și el destul de laborios.

O verificare simplă, dar laborioasă a normalității distribuției datelor, o reprezintă aprecierea empirică a similarității dintre poligonul frecvențelor valorilor variabilei investigate și curba în formă de clopot a distribuției normale. În cadrul exercițiului din exemplul 2.3 (fig. 2.3, 2.4) se poate spune că frecvențele observate au o distribuție ce poate fi considerată aproximativ normală.

Verificarea normalității datelor se poate face și pornind de la proprietățile matematice ale distribuției normale. Astfel, dacă valorile sunt distribuite simetric față de medie, adică între medie și mediană nu există o diferență mare (secțiunea 3.1 – Relația dintre descriptorii tendinței centrale), și aproximativ 70% din valori sunt cuprinse în intervalul delimitat de valorile $(\bar{x} - s)$ și $(\bar{x} + s)$, atunci se poate aprecia că variabila analizată este aproximativ normal distribuită în probă.

Exemplul 4.9. Să se verifice rapid dacă datele din exemplul 3.1 au o distribuție apropiată de cea normală.

Simpla vizualizare a descriptorilor tendinței centrale arată că distribuția valorilor din probă este aproximativ normală deoarece aceștia au valori foarte apropiate.

$Mo = 69$
$Me = 69{,}8$
$\bar{x} = 70{,}85$

Intervalul $\bar{x} \pm s$ are următoarele limite:

$70{,}85 - 6{,}92 = 63{,}93$
$70{,}85 + 6{,}92 = 77{,}77$.

Între aceste două valori se găsesc 34 din 50 de valori, adică 68% din valorile din probă. Acest procent este apropiat de cel cuprins în intervalul $\mu - \sigma, \mu + \sigma$ al unei distribuții normale, adică de 68,26%.

Dacă datele nu sunt normal distribuite, atunci cel mai simplu este să

se folosească o metodă statistică alternativă, neparametrică. Folosirea metodelor parametrice în astfel de situații este totuși permisă dacă se realizează o **transformare a datelor** care să corecteze distribuția acestora.

Transformarea datelor este necesară dacă datele sunt sub formă de număr de entități. Astfel de variabile discrete au o distribuție evident asimetrică. În astfel de situații se folosesc transformări care au rolul de a **normaliza** distribuția datelor.

Numeroase tehnici parametrice compară mediile probelor care se presupune că au varianțe suficient de asemănătoare și care, din această cauză, pot fi ignorate. Datele discrete ce reprezintă numărători de entități nu îndeplinesc această condiție, deoarece varianța este dependentă de medie în sensul că populațiile la care media are o valoare mare, împrăștierea valorilor față de medie este mai mare și, implicit, varianța va fi mai mare. În astfel de situații transformările au rolul de a întrerupe relația dintre medie și varianță, adică de a **stabiliza varianța** datelor.

Transformările cele mai utilizate în ecologie sunt logaritmul, radicalul și transformarea arcsin (anexa 3). Acestea se calculează în diferite condiții pentru toate valorile individuale din probe.

Transformarea logaritmică se utilizează atunci când varianța probei este mai mare decât media acesteia. De asemenea, are si rolul de a normaliza distribuția datelor. Valoarea transformată x' a unei valori individuale x se calculează folosind cel mai adesea logaritmul zecimal sau pe cel natural:

$$x' = log_{10}(x) \qquad \text{sau} \qquad x' = ln(x)\,.$$

Dacă în probă există valori egale cu zero, atunci logaritmul nu are sens și transformarea se poate face fie adăugând o unitate la toate valorile, fie folosind transformarea arcsinh (sinus hiperbolic invers) returnează valoarea zero dacă $x = 0$.

$$x' = log(x + 1)$$
$$x' = arcsinh(x)$$

Transformarea prin extragerea radicalului se folosește atunci când varianța datelor de tip număr de entități este aproape egală cu media și

pentru normalizarea distribuției. Și aici, în cazul existenței valorilor nule, se poate aduna o constantă la toate valorile din probă (1 sau 0,5):

$$x' = \sqrt{x}$$
$$x' = \sqrt{x+1} \quad \text{sau} \quad x' = \sqrt{x+0{,}5}\,.$$

Transformarea arcsin se folosește atunci când datele sunt sub formă de proporții sau procente. În cazul unor astfel de variabile distribuția valorilor este trunchiată de cele două valori extreme: 0 și 1 pentru proporții și 0 și 100 pentru procente. Rezultatul transformării în radian se transformă în grade (anexa 3).

$x' = arcsin(\sqrt{x})$ pentru proporții

$x' = arcsin\left(\sqrt{\frac{x}{100}}\right)$ pentru procente

În anumite situații este nevoie de realizarea transformării inverse (anexa 3) pentru raportarea rezultatului în forma inițială a datelor ($x' \rightarrow x$).

Transformarea inversă se realizează astfel:

Pentru transformarea prin logaritmare:

$x = antilog(x')$

$x = antilog(x') - 1\,.$

Pentru transformarea arcsinh:

$s = sinh(x')\,.$

Pentru transformarea prin radical:

$x = x'^2$

$x = x'^2 - 1 \quad \text{sau} \quad x = x'^2 - 0{,}5\,.$

Pentru transformarea arcsin:

$x = (\sin(x'))^2$

$x = (\sin(x'))^2 \cdot 100\,.$

5. STATISTICĂ INFERENȚIALĂ: ELEMENTE INTRODUCTIVE

Statistica inferențială (inductivă sau analitică) este partea statisticii care cuprinde metode de apreciere critică a variabilității parametrilor empirici. Inferența statistică reprezintă tratarea teoretică a datelor pentru a se ajunge la concluzii logice, asociate observațiilor efectuate. Din punct de vedere ecologic, inferența statistică reprezintă stabilirea unor concluzii despre populații pornindu-se de la analiza probelor prelevate din populațiile respective.

În general, se recunosc două categorii largi de inferențe statistice: estimarea unor parametri populaționali și testarea ipotezelor statistice.

5.1. ESTIMAREA MEDIEI POPULAȚIONALE

Dacă dintr-o populație se prelevează o probă aleatoare, aceasta va fi una din numeroasele probe aleatoare care se pot extrage din populația respectivă. Fiecare dintre aceste populații va avea statistici diferite: medii diferite, deviații standard diferite. Cu toate acestea, statisticile acestor populații sunt estimatori ai parametrilor populaționali (fig. 5.1). Diferențele dintre aceste probe sunt cauze ale erorii de eșantionare, ce rezultă din faptul că unele probe vor cuprinde mai multe valori mari, în timp ce altele, mai multe valori mici din populația de cercetat. Eroarea de eșantionare nu este rezultatul unor greșeli realizate de observator, ci reflectă împrăștierea aleatoare ce se regăsește în probe. Mediile probelor prelevate aleator din populație se distribuie în jurul mediei populaționale, la fel cum valorile individuale într-o probă se distribuie în jurul mediei probei (fig. 5.2). Acest concept are o valoare fundamentală și este surprins de **Teorema limită centrală** (Moivre 1738, Laplace 1813): mediile probelor ($\bar{x}$) extrase dintr-o populație normal distribuită sunt la rândul lor normal distribuite în jurul mediei populaționale (μ). Mediile probelor extrase dintr-o populație nenormal distribuită au o distribuție aproximativ normală dacă dimensiunea probei este mare ($n > 30$).

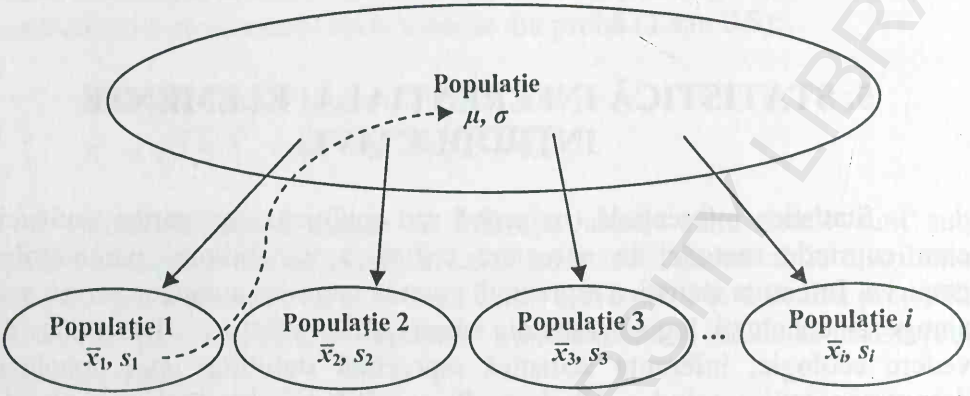


Figura 5.1. Reprezentarea grafică a prelevării repetate a probelor din populație. (linie continuă – sensul prelevării; line întreruptă – sensul estimării)

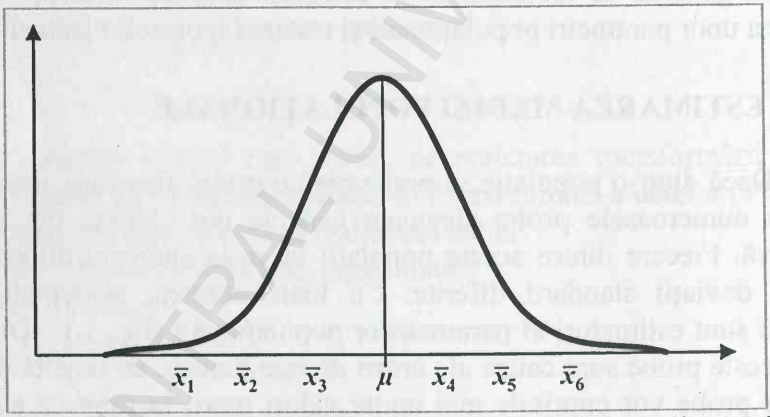


Figura 5.2. Distribuția normală a mediilor probelor față de media populației

Utilitatea acestei teoreme constă în faptul că nu este necesară prelevarea repetată a probelor din populație pentru a cunoaște modul lor de distribuire; ele vor avea o distribuție normală. Astfel, putem lua în considerare doar o singură medie a unei probe prelevate dintr-o populație ca fiind una dintre numeroasele medii a căror distribuție ar fi normală. La fel cum deviația standard surprinde împrăștierea valorilor individuale față de media probei, tot așa, împrăștierea mediilor probelor poate fi surprinsă de o deviație standard a mediilor probelor, care se numește **eroarea standard a mediei**.

Estimarea mediei populaționale se poate face pornind de la media și deviația standard ale unei probe și cu ajutorul erorii standard a mediei. Dat fiind faptul că distribuția mediilor probelor se abate de la normalitate pe măsură ce dimensiunea probei scade, se apelează la o distribuție care descrie mai bine distribuția mediilor probelor atunci când deviația standard a populației este estimată prin deviația standard a probei. Această distribuție se numește **distribuția *t* sau distribuția Student**.

Distribuția Student este similară în multe privințe cu distribuția normală, dar, spre deosebire de aceasta, este definită, pe lângă media și deviația standard, și de numărul gradelor de libertate ($n-1$). Așa cum o valoare z corespunde unei anumite proporții din distribuția normală standard, tot așa o valoare t corespunde unei proporții a distribuției Student, dar în plus ia în considerație și dimensiunea probei prin intermediul gradelor de libertate. Valorile lui t scad odată cu creșterea diferenței $n-1$, astfel că o valoare critică $t_{(0,05,\infty)}$ ce definește 0,95 sau exclude 0,05 din distribuția Student pentru o infinitate de valori ca grade de libertate are valoarea 1,96, adică exact valoarea lui $z_{(0,95)}$ ce definește aceeași proporție din distribuția normală standard. Deci distribuția t tinde să devină normală odată cu creșterea dimensiunii probei.

Valorile distribuției t sunt tabelate sau se pot calcula în funcție de proporția exclusă din distribuție și de numărul gradelor de libertate. De exemplu, valoare t ce exclude 0,05 din distribuția Student pentru $n-1=4$ grade de libertate este 2,776. Proporția (0,05) sau procentul (5%) exclus este repartizat în mod egal în cele două cozi ale distribuției (fig. 5.3).

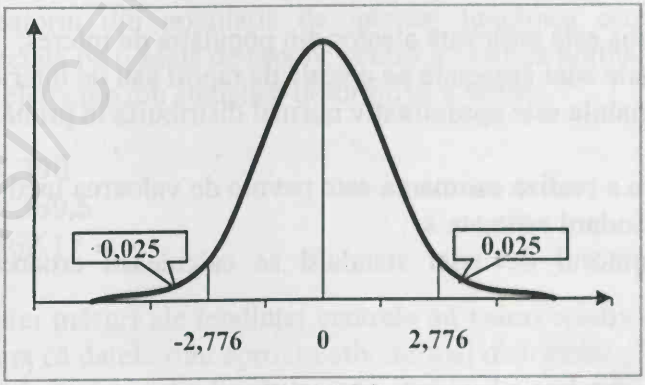


Figura 5.3

Deci există de fapt două valori t: $+2{,}776$ care exclude $0{,}05/2 = 0{,}025$ din coada din dreapta a distribuției și $-2{,}776$ care exclude $0{,}025$ din coada stângă a distribuției.

Dacă dorim să reprezentăm valoarea t care exclude $0{,}05$ din distribuție doar în coada dreaptă pentru 4 grade de libertate, atunci trebuie căutată în tabel valoarea ce exclude $0{,}1$ din distribuție, care exclude câte $0{,}05$ în fiecare coadă din distribuție (fig. 5.4). Această valoare este $\pm 2{,}132$. Deci, dacă ne interesează o singură coadă a distribuției, trebuie căutată valoarea t care exclude o proporție dublă din distribuție.

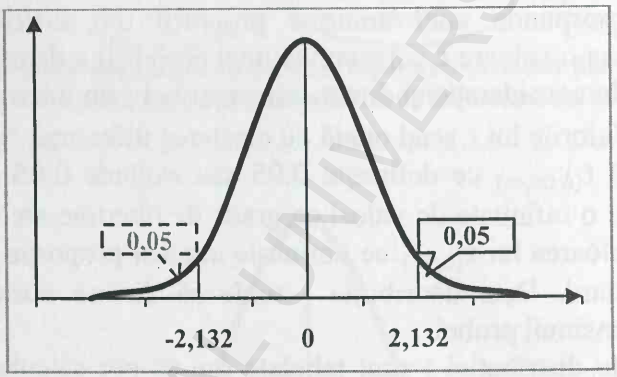


Figura 5.4

Estimarea intervalului de confidență al mediei populaționale pe baza deviației standard a probei și cu ajutorul distribuției t are următoarele condiții de aplicare:

1. proba este prelevată aleator din populația de interes;
2. datele sunt apreciate pe o scală de raport sau de interval;
3. variabila este aproximativ normal distribuită în probă.

Pentru a realiza estimarea este nevoie de valoarea mediei probei $\bar{x}$ și a deviației standard estimate s.

Cu ajutorul deviației standard se calculează eroarea standard a mediei:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

Se estimează intervalul de confidență pentru o probabilitate de 0,95 (95%) a mediei populației pornind de la relația:

$$\mu = \bar{x} \pm s_{\bar{x}} \cdot t_{(0,05,n-1)} \cdot$$

Din această relație rezultă limita inferioară (LI) și cea superioară (LS) a intervalului de confidență:

$$LI = \bar{x} - s_{\bar{x}} \cdot t_{(0,05,n-1)}$$
$$LS = \bar{x} + s_{\bar{x}} \cdot t_{(0,05,n-1)} \cdot$$

Concluzia estimării este că intervalul LI-LS include media populației din care a fost extrasă proba, cu o probabilitate de 95% (0,95).

Exemplul 5.1. La o probă formată din 30 de indivizi de viperă de stepă (*Vipera ursinii moldavica*) extrasă aleatoriu dintr-o populație s-a măsurat lungimea în mm de la vârful botului și până la cloacă. S-a estimat apoi intervalul de confidență al mediei pentru o probabilitate de 95%.

390	228	440	66	215	146	443	375	450	260
330	500	340	363	491	325	390	425	418	422
389	435	470	360	370	400	390	430	164	340

Se verifică dacă datele îndeplinesc condițiile de aplicare: proba a fost prelevată aleatoriu din populația de interes; lungimea este o variabilă continuă apreciată pe o scală de raport; pentru a verifica normalitatea datelor putem folosi elemente din statistica descriptivă a probei.

$$Mo = 390$$
$$Me = 389,5$$
$$\bar{x} = 362,17$$

Cele trei măsuri ale tendinței centrale au valori relativ apropiate. Se poate considera că datele sunt aproximativ normal distribuite.

În continuare se calculează deviația standard a probei:

$s = 96{,}62$.

Cu ajutorul acestei valori se află eroarea standard a mediei:

$$s_{\bar{x}} = \frac{96{,}62}{\sqrt{30}} = \frac{96{,}62}{5{,}47} = 17{,}64 .$$

Se estimează limitele intervalului de confidență al mediei. În acest sens, se caută valoarea t în tabel.

$$t_{(0{,}05,29)} = 2{,}045$$

$$\mu = 362{,}17 \pm 17{,}64 \cdot 2{,}045 = 362{,}17 \pm 36{,}08$$
$$LI = 362{,}17 - 36{,}08 = 326{,}09$$
$$LS = 362{,}17 + 36{,}08 = 398{,}25$$

Intervalul 326,09-398,25 include media populațională a lungimii de la vârful botului la cloacă cu o probabilitate de 95%.

Reprezentarea grafică a intervalului de confidență al mediei se realizează prin intermediul unor segmente dispuse deasupra și dedesubtul mediei, ce simbolizează limita inferioară și cea superioară (fig. 5.5).

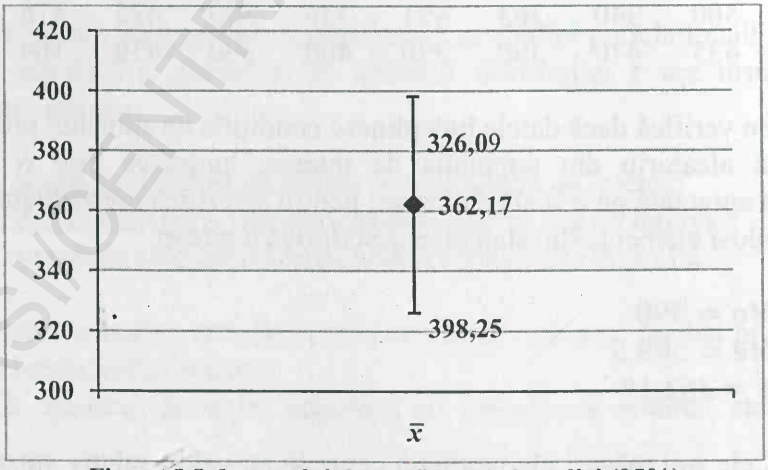


Figura 5.5. Intervalul de confidență al mediei (95%)

Acest tip de reprezentare poate fi folosit și atunci când se compară

mediile mai multor probe din populații diferite. Suprapunerea intervalelor indică absența unei diferențe marcante între mediile celor două populații din care au fost prelevate probele analizate.

5.2. ESTIMAREA UNEI PROPORȚII

În ecologie, se folosesc adesea frecvențele de apariție ale unei valori în probe reprezentată sub forma unei proporții sau procent din total. Proporția unei specii în probă reprezintă o estimare a proporției populației în comunitatea analizată. Probele ulterioare reprezintă estimatori ai proporției populației din specia de interes în comunitate. Proporțiile rezultate din analiza acestor probe vor fi diferite datorită erorii de selecție. Aceste proporții ale probelor se vor distribui în jurul proporției populației în același fel cu modul în care se distribuie mediile probelor în jurul mediei populației. Deviația standard a distribuției se numește eroare standard și se estimează astfel:

$$E.S. = \sqrt{\frac{p(1-p)}{n-1}}$$

p – proporția speciei n – numărul total de specii.

Intervalul de confidență al proporției populației este:

$$p \pm (1{,}96 \cdot E.S.)\,.$$

5.3. ESTIMAREA EFECTIVULUI POPULAȚIONAL

Indicele Lincon-Petersen este un estimator al numărului de indivizi dintr-o populație, pe baza proporției indivizilor marcați în prima probă care se regăsesc (sunt recapturați) în a doua probă.

$$\widehat{N} = \frac{(n_1+1)(n_2+1)}{(m_2+1)} - 1\,,$$

$\widehat{N}$ – estimarea efectivului populației;
n_1 – nr. indivizilor capturați, marcați și eliberați în prima probă;
n_2 – nr. total al indivizilor capturați în a doua probă;
m_2 – nr. indivizilor marcați, recapturați în a doua probă.

Deviația standard aproximativă a acestui estimator este:

$$s = \sqrt{\frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}}.$$

Intervalul de confidență al efectivului populației (N) pornind de la relația:

$$N = \widehat{N} \pm 1{,}96 \cdot s\,.$$

5.4. ESTIMAREA INDICELUI DE DIVERSITATE

Indicii de diversitate sunt utili în aprecierea biodiversității unei zone. Cel mai simplu indicator al diversității biologice este bogăția specifică sau numărul de specii. Există o serie de indici ai diversității care, pe lângă numărul de specii, iau în considerație și cât de echitabil sunt reprezentate speciile din comunitate, prin intermediul numărului de indivizi. Unul dintre cei mai folosiți astfel de indici este **indicele Shannon-Weaver**:

$$H = -\sum p_i \cdot ln(p_i)$$

p_i – proporția indivizilor speciei i din suma nr. de indivizi ai tuturor speciilor.

Estimarea deviației standard care să descrie împrăștierea valorilor indicilor calculați pentru aceeași comunitate, în jurul unei medii populaționale, este dificilă. Din rațiuni practice, valoarea indicelui trebuie tratată ca o variabilă ordinală. Astfel, o valoare a indicelui egală cu 4,0 nu trebuie considerată ca fiind de două ori mai mare decât una egală cu 2,0. Tehnicile statistice care se pot aplica valorilor apreciate pe o scală ordinală sunt în general neparametrice sau independente de distribuție. De exemplu, un set de indici obținuți prin analiza mai multor probe extrase din aceeași zonă poate fi comparat cu un alt set extras rezultat din altă zonă prin intermediul testului U Mann-Whitney (secțiunea 7.1.2).

O altă modalitate de comparare a indicilor constă în transformarea acestora în diversități relative (H_{rel}), exprimate proporțional sau procentual:

$H_{rel} = \frac{H}{H_{max}} = \frac{H}{ln(S)}$

H_{max} – diversitatea maximă pentru același număr de specii (diversitatea unei comunități ideale cu același nr. de specii cu cea reală, în care toate speciile sunt reprezentate prin același nr. de indivizi);

S – bogăția specifică sau nr. de specii identificate în comunitate.

Când se compară valorile indicilor de diversitate (H) trebuie avute în vedere două aspecte: se compară indici pentru comunități asemănătoare (de exemplu, se compară o comunitate de păsări cu alta tot de păsări, nu de mamifere); se compară indici rezultați din analiza unor probe cu numere apropiate de organisme.

Exemplul 5.2. Într-un studiu al vegetației de stepă din rezervația de la Valea lui David s-au calculat diversitățile pentru cinci comunități vegetale din asociația *Taraxaco serotinae-Festucetum valesiacae* și diversitatea unei comunități din aceeași asociație, studiată înainte de 1969 și notată cu A.

Releveu	H	H_{rel}	E	S	$ln(E)$	$ln(S)$
1	1,519	0,438	0,143	32,0	-1,947	3,466
2	1,543	0,454	0,156	30,0	-1,858	3,401
3	1,971	0,579	0,239	30,0	-1,430	3,401
5	2,051	0,637	0,311	25,0	-1,168	3,219
4	2,130	0,662	0,337	25,0	-1,089	3,219
Media	1,843	0,554	0,237	28,4	-1,498	3,341
A	1,639	0,423	0,107	48,0	-2,232	3,871

Se poate observa că media indicilor de diversitate a celor cinci relevee actuale este ceva mai mare decât valoarea indicelui pentru releveul A, deși numărul de specii (S) în cel din urmă este evident mai mare. Acest fapt poate fi explicat prin echitabilitatea redusă a speciilor în releveul A.

Echitabilitatea în acest exemplu a fost calculată pe baza relației:

$$E = \frac{e^H}{S} => H = ln(S) + ln(E) .$$

Cum echitabilitatea este un număr subunitar, logaritmul va fi un număr negativ, deci indicele de diversitate H este egal cu diversitatea maximă $H_{max} = ln(S)$ care este micşorată de echitabilitatea scăzută a abundenţei speciilor $ln(E)$

5.5. TESTAREA IPOTEZELOR STATISTICE

În orice ştiinţă, progresul se obţine prin realizarea observaţiilor asupra unor procese sau fenomene şi prin experimente ale căror concluzii sunt utilizate sub forma unor generalizări sau teorii care să explice observaţiile. Demersul ştiinţific debutează cu realizarea observaţiilor şi cu explicarea lor. Explicaţia unei observaţii ştiinţifice se numeşte **ipoteză** şi are următoarele caracteristici: este în concordanţă cu observaţiile făcute, adică, dacă este adevărată, atunci va explica ceea ce s-a observat; poate fi testată prin experimente, adică, dacă este falsă, atunci acest lucru poate fi dovedit.

De ce trebuie dovedită falsitatea unei ipoteze şi nu veridicitatea ei? În filosofia ştiinţei, se consideră că să poate dovedi că o ipoteză falsă este falsă, în timp ce o ipoteză adevărată poate să nu se dovedească niciodată că este adevărată. Ca urmare, o ipoteză este considerată adevărată atât timp cât nu poate fi infirmată prin alte observaţii, experimente şi testări. Când încercările de a dovedi falsitatea unei ipoteze eşuează, atunci încrederea, confidenţa în ipoteza respectivă creşte. Dacă o astfel de ipoteză are o aplicativitate largă şi explică numeroase evenimente, atunci ea devine o **teorie**. La fel ca în cazul ipotezelor, o teorie adevărată s-ar putea să nu poată fi dovedită a fi adevărată, în timp ce una falsă se poate dovedi a fi falsă.

Se poate astfel spune că ştiinţa avansează mai degrabă infirmând decât afirmând şi că până la urmă teoriile incorecte vor fi invalidate.

Metodologia ştiinţifică de confirmare a unei ipoteze operează pe baza logicii „dacă atunci": **dacă** ipoteza este corectă, **atunci** rezultatul testării trebuie să fie unul anume. Dacă rezultatul testării este altul decât cel prezis de ipoteză, aceasta se respinge şi trebuie căutată o explicaţie mai bună. Acest proces tipic pentru ştiinţă este numit **testarea ipotezelor**.

Testarea unor concluzii ştiinţifice prin procedee statistice se numeşte **testarea ipotezelor statistice** şi reprezintă o aplicare specifică a metodologiei ştiinţifice. Formularea ipotezelor statistice se face astfel încât

să existe doar două rezultate posibile. De exemplu, se pot formula două ipoteze: „afirmația A este adevărată" și „afirmația A nu este adevărată". Dacă primul enunț este cel adevărat, atunci, conform filosofiei științei, nu se poate dovedi acest fapt. Dacă însă se testează al doilea enunț și acesta se dovedește a fi incorect (o ipoteză falsă se poate dovedi că este falsă), atunci se respinge enunțul testat – „afirmația A este falsă" – și se acceptă celălalt enunț – „afirmația A este adevărată" – ca unică alternativă corectă.

Când se lucrează pe probe extrase din populații, deci doar cu o parte din întregul la care în final se va face referință, va exista întotdeauna o probabilitate ca proba să nu fie reprezentativă pentru toată populația. Cu toate acestea, se va putea preciza probabilitatea ca o ipoteză din cele două să fie corectă sau incorectă. Dacă probabilitatea ca ea să fie incorectă este foarte mică, atunci se poate considera că ipoteza respectivă este corectă și invers, dacă probabilitatea ca ipoteza să fie corectă este foarte mică, atunci se poate concluziona că ipoteza este incorectă.

În orice testare a ipotezelor statistice, ipotezele formulate sunt întotdeauna contradictorii. Ipoteza testată prin diferite procedee numite **teste statistice** este așa-numita **ipoteză nulă** (H_0). Aceasta presupune în general lipsa efectului, lipsa diferenței și, ca urmare, conține sau implică o egalitate. Cealaltă ipoteză, **ipoteza alternativă** (H_1 sau H_a), se numește ipoteză alternativă. De exemplu, dacă dorim să arătăm că A este diferit de B, atunci H_0 va fi $A = B$ (conține o egalitate), iar cea H_1 va fi $A \neq B$. Ipoteza care se testează este de fapt H_0. Dacă aceasta se dovedește adevărată, atunci se acceptă ca atare. Dacă H_0 se dovedește a fi falsă, atunci se respinge și se acceptă H_1 ca unică alternativă.

Orice test statistic constă într-o serie de calcule aplicate datelor din probe care au ca rezultat o singură valoare numită **statistica testului**. Statistica unui test reprezintă o translație a datelor din probe la o distribuție cunoscută. Este un proces similar cu cel de trecere a unei valori de la o distribuție normală particulară la o valoare z a distribuției normale standard, valoare ce corespunde unei anumite proporții din distribuție sau unei anumite probabilități (secțiunea 4), conform schemei $x \rightarrow z \rightarrow p$. Statistica testului, proprie unui anumit tip de distribuție, este comparată cu o valoare cu semnificație de prag pentru o anumită probabilitate, numită **valoare critică**. În funcție de poziționarea statisticii testului față de valoarea critică, se ia decizia de acceptare sau respingere a ipotezei nule. Valorile

critice pentru fiecare test statistic sunt calculate și aranjate în tabele sau se pot calcula pornind de la funcțiile specifice distribuțiilor.

În funcție de întrebarea la care trebuie să răspundă testul statistic, ipotezele acestuia se pot scrie în mai multe variante. Dacă ipotezele conțin semnele $=$ și $\neq$, atunci se realizează un **test în variantă bilaterală** ($H_0: A = B$; $H_1: A \neq B$). Denumirea provine de la faptul că există două situații în care se poate respinge ipoteza nulă și accepta ipoteza alternativă: când $A > B$ și când $A < B$. Dacă ipotezele conțin semnele $\geq$,$\leq$,$>$ și $<$, atunci se realizează un **test în variantă unilaterală** ($H_0: A \leq B; H_1: A > B$ sau $H_0: A \geq B; H_1: A < B$). În oricare din cele două variante există doar o singură situație în care se poate respinge ipoteza nulă și se poate accepta ipoteza alternativă: dacă și numai dacă $A > B$, pentru prima pereche de ipoteze, și dacă și numai dacă $A < B$, pentru a doua pereche de ipoteze.

În general, testele unilaterale se utilizează doar dacă există un motiv apriori care să sugereze o tendință direcțională a datelor. Este bine ca testele bilaterale să se facă după o testare în variantă bilaterală. Între cele două variante ale unui test nu există nici o diferență în privința modului de calcul al statisticii testului, ci diferă doar ipotezele și pragul de semnificație mai mic în cazul testelor unilaterale (secțiunea 5.1, fig. 5.3, 5.4 și explicațiile aferente).

Luarea unei decizii statistice se realizează în funcție de pragul de probabilitate. Acesta se mai numește și **nivel de confidență**, **de încredere** sau **prag de semnificație** și se notează cu α. Valorile α cel mai des utilizate în ecologie sunt 0,05 sau 0,01 și se desemnează înainte de derularea testului. Pragul de semnificație (α) sau probabilitatea calculată pentru o anumită statistică a unui test (p) trebuie precizată în concluziile oricărei cercetări în care s-au folosit teste statistice (de exemplu, „rezultatul este semnificativ în proporție de 95%" sau „ ... pentru $\alpha = 0{,}05$" sau „ ... pentru $p < 0{,}05$" sau „ ... pentru $p = 0{,}0003$").

Se poate întâmpla ca H_0 să fie respinsă pentru o valoare a probabilității egale cu 0,05, dar să nu poată fi respinsă dacă nivelul de încredere stabilit apriori este de 0,01. Această situație se datorează faptului că, pentru majoritatea distribuțiilor, valoarea critică crește pe măsură ce nivelul de încredere scade. Ce decizie se ia într-o astfel de situație și cum poate fi ea argumentată pentru a elimina subiectivismul?

În orice test statistic pot să apară două genuri de **erori statistice** (tab. 5.1):

– **eroare de genul I**, ce constă în respingerea eronată a H_0 când este adevărată; riscul sau probabilitatea de a face o astfel de eroare este α;

– **eroare de genul II**, ce constă în acceptarea eronată a H_0 când este falsă; riscul sau probabilitatea de a face o astfel de eroare este β.

Tabelul 5.1. Consecințele posibile ale unei decizii statistice

		Ipoteza adevărată	
		H_0	H_1
Ipoteza acceptată	H_0	Corect $(1-\alpha)$	Eroare II $p=\beta$
	H_1	Eroare I $p=\alpha$	Corect $(1-\beta)$

Dacă se dorește reducerea riscului de a comite o eroare I, atunci α trebuie să scadă, ceea ce conduce la creșterea riscului β de a comite o eroare II și invers. Se consideră că $\alpha = 0{,}05$ asigură un echilibru între riscul de a comite o eroare de genul I și cel de a comite o eroare de genul II.

Dacă valoarea α se decide de la început sau se poate calcula pentru o anumită statistică a unui test corespunzător unei funcții de distribuție, valoarea lui β nu se calculează. Valoarea lui β scade pe măsură ce dimensiunea probelor (n) crește și crește pe măsură ce diferența dintre valorile comparate (A și B) scade. Riscul β variază de la un test la altul. Un **test puternic** înseamnă de fapt că are un **risc β mic**, adică este mai puțin influențat de dimensiunea probei și de diferențele mici dintre valorile comparate.

Legat de puterea unui test sau de riscul de a comite o eroare de genul II, trebuie menționat că testele neparametrice sau independente de distribuție sunt mai puțin puternice decât cele parametrice, mai restrictive. Din această cauză un cercetător ar putea manifesta o tendință de evitare a testelor neparametrice în ideea folosirii unor teste mai puternice. O astfel de atitudine se poate dovedi eronată – nu trebuie sacrificată validitatea utilizării unui test în favoarea puterii acestuia! Regula de siguranță în privința alegerii unui test parametric sau neparametric este că, dacă există o îndoială oricât de mică cu privire la modul în care datele din probe satisfac condițiile

restrictive ale unui anumit test parametric, atunci mai bine se apelează la un test neparametric alternativ celui parametric.

Rezumând aspectele prezentate până acum, **testarea ipotezelor statistice** se realizează prin parcurgerea următoarelor **etape**:

1. Enunțarea clară a întrebării la care se dorește aflarea răspunsului în urma prelucrării datelor din probe.
2. Identificarea tipului de variabilă și a scalei de apreciere a acesteia și aprecierea distribuției probei. Această etapă permite luarea deciziei privind utilizarea unui test parametric sau a unuia neparametric.
3. Pe baza răspunsurilor din primele două etape se formulează cele două ipoteze statistice (practic se alege o variantă bilaterală sau una unilaterală a testului) și se stabilește regula de decizie (se desemnează nivelul de încredere sau de confidență α).
4. Se calculează statistica sau statisticile testului.
5. Se compară statistica obținută cu valoarea critică corespunzătoare valorii α și gradelor de libertate și se ia o decizie privind acceptarea sau respingerea ipotezei nule. Decizia mai poate fi luată și prin calcularea probabilității corespunzătoare statisticii testului (aceasta va fi de fapt probabilitatea ca H_0 să fie adevărată) cu ajutorul funcției distribuției acesteia sau folosind un sistem de programe pentru computere adecvat (un software).

6. TESTAREA UNEI IPOTEZE PRIVIND MEDIA UNEI SINGURE POPULAȚII

Această testare permite compararea mediei unei probe ($\bar{x}$) cu o valoare de interes care de obicei reprezintă media cunoscută a unei populații (μ). Altfel spus, se verifică ce probabilitate există ca proba luată în analiză să provină dintr-o populație cu o anumită medie cunoscută. Populația din care a fost extrasă proba poate fi diferită de cea de referință, caz în care se testează o ipoteză nulă conform căreia nu există o diferență semnificativă între mediile celor două populații.

Testul care se folosește într-o astfel de situație se numește **Testul *t* (Student) pentru o probă**. Fiind un test parametric, condițiile de aplicare ale acestuia sunt următoarele:

1. proba trebuie să fie extrasă aleator din populație;
2. variabila trebuie să fie exprimată pe o scală de raport sau de interval;
3. valorile probei trebuie să fie aproximativ normal distribuite.

Dacă μ este media populației din care a fost extrasă proba și μ_0 este media populației de referință sau o valoare de referință, atunci ipotezele testului pot fi:

$$H_0: \mu = \mu_0 \quad H_0: \mu \leq \mu_0 \quad H_0: \mu \geq \mu_0$$
$$H_1: \mu \neq \mu_0 \quad H_1: \mu > \mu_0 \quad H_1: \mu < \mu_0 .$$

Prima pereche de ipoteze se scrie în cazul variantei bilaterale a testului, adică atunci când întrebarea este: „Există o diferență semnificativă între μ și μ_0?".

Ultimele două perechi de ipoteze se scriu în cazul unui **test unilateral dreapta** („Este μ semnificativ mai mare decât μ_0?") și, respectiv, unui **test unilateral stânga** („Este μ semnificativ mai mică decât μ_0?").

Indiferent de varianta în care se realizează testul, statistica sa este:

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}.$$

În această relație $\bar{x}$ este un estimator al mediei populației din care a fost extrasă proba (μ).

Condiția testului constă în compararea statisticii acestuia cu o valoare critică $t_{(\alpha,n-1)}$:

dacă $t \geq t_{(\alpha,n-1)} \Rightarrow H_0$ se respinge și se acceptă H_1 pentru o probabilitate de $1-\alpha$ sau $100(1-\alpha)\%$. Deci se acceptă că μ și μ_0 diferă semnificativ una de alta. Dacă $t < t_{(\alpha,n-1)}$, atunci se acceptă H_0 pentru aceeași probabilitate, adică nu există o diferență semnificativă între μ și μ_0.

Dacă μ este semnificativ diferită de μ_0, înseamnă că μ este ori mai mare, ori mai mică decât μ_0 Aceasta implică faptul că $\bar{x}$ este mai mic sau mai mare decât μ_0.

În situația în care $\bar{x}$ este mai mare decât μ_0, atunci numărătorul statisticii t va fi negativ și, implicit, statistica t va avea o valoare negativă. Deci t trebuie comparat cu valoarea critică pozitivă aflată în dreapta cozii distribuției t. În același mod, dacă $\bar{x}$ este mai mic decât μ_0, statistica t va fi negativă și trebuie comparată cu valoarea critică negativă din coada stângă a cozii distribuției.

Condiția testului ar trebui în realitate scrisă astfel:

dacă t este mai mic decât valoarea critică negativă și mai mare decât valoarea critică pozitivă, atunci H_0 se respinge și se acceptă H_1 pentru o probabilitate de $1-\alpha$ sau $100(1-\alpha)\%$.

Astfel rezultă că dacă t se găsește între valoarea critică negativă și cea pozitivă, H_0 va fi acceptată. Deci între cele două valori critice există zona de acceptare a ipotezei nule ($1-\alpha$), în afara căreia se găsesc zonele de respingere a acesteia și de acceptare a ipotezei alternative (câte $\alpha/2$ în fiecare coadă a distribuției) (fig. 6.1).

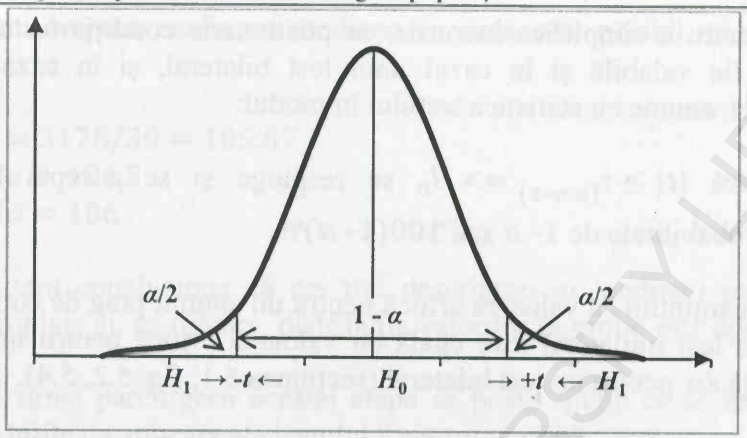


Figura 6.1. Zonele de respingere a H_0 pentru un test bilateral

Atunci când se aplică testul în variantă unilaterală, atunci există o singură zonă de respingere a H_0 cantonată doar într-o singură coadă a distribuției: dreaptă sau stângă. Dacă se vizează coada din dreapta distribuției ($H_1: \mu > \mu_0$), atunci statistica testului (t) trebuie comparată cu valoarea critică pozitivă (fig. 6.2).

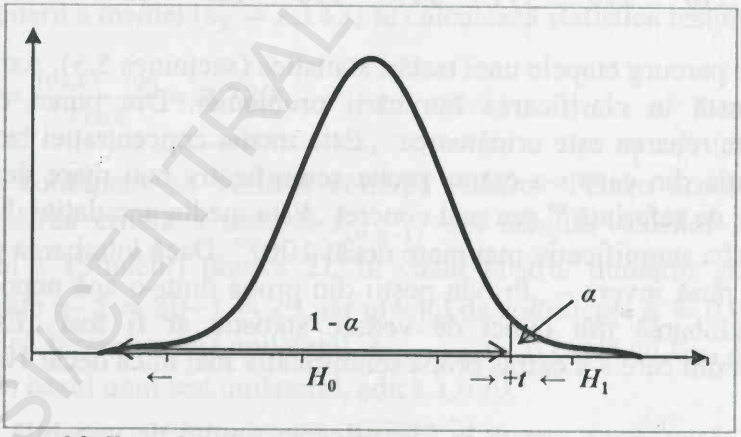


Figura 6.2. Zona de respingere a H_0 pentru un test unilateral dreapta

Dacă se urmărește coada din stânga a distribuției ($H_1: \mu < \mu_0$), atunci statistica testului (t) trebuie comparată cu valoarea critică negativă (zona de respingere a H_0 se află în partea opusă față de cum este prezentată în figura 6.2).

Pentru a simplifica lucrurile, se poate scrie condiția testului astfel încât să fie valabilă și în cazul unui test bilateral, și în cazul testelor unilaterale, anume cu statistica testului în modul:

dacă $|t| \geq t_{(\alpha,n-1)} => H_0$ se respinge și se acceptă H_1 cu o probabilitate de $1-\alpha$ sau $100(1-\alpha)\%$.

Reamintim că valoarea critică pentru un anumit prag de confidență α pentru un test unilateral este egală cu valoarea critică pentru un prag de confidență 2α pentru un test bilateral (secțiunea 5.1, fig. 5.2, 5.4).

Exemplul 6.1. Într-un studiu s-a urmărit concentrația unui biomarker al poluării apei în corpul unei specii de pește. O concentrație mai mare de 100 unități/g indică o poluare a apei în care trăiesc peștii. Este poluată apa din care s-a extras aleatoriu o probă formată din 30 de pești?

87	90	94	94	94	95	95	98	98	101
101	102	102	103	104	105	106	106	106	107
108	110	110	111	117	118	123	124	130	137

Se parcurg etapele unei testări statistice (secțiunea 5.5). Astfel, prima etapă constă în clarificarea întrebării problemei. Din punct de vedere statistic, întrebarea este următoarea: „Este media concentrației biomarker-ului în populația din care s-a extras proba semnificativ mai mare decât media populației de referință?" sau mai concret „Este media populației din care s-a extras proba semnificativ mai mare decât 100?". Dacă întrebarea problemei ar fi fost pusă invers – „Provin peștii din probă dintr-o apă nepoluată?" –, atunci întrebarea din punct de vedere statistic ar fi fost „Este media populației din care s-a extras proba semnificativ mai mică decât 100?".

A doua etapă constă în identificarea tipului de variabilă. Așa cum reiese din textul problemei, variabila este concentrația biomarker-ului pe gram. Concentrația este apreciată pe o scală de raport, deoarece valoarea zero este absolută. Un alt aspect ce trebuie urmărit în această etapă este distribuția valorilor. Pentru a aprecia distribuția ne folosim de relația dintre măsurile tendinței centrale într-o probă normal distribuită (secțiunea 3.1).

Deci vom compara media cu mediana și, dacă este posibil, și cu modul probei:

$\bar{x} = 3176/30 = 105{,}87$
$Me = 104{,}5$
$Mo = 106$.

Putem concluziona că cei trei descriptori ai tendinței centrale au valori apropiate și, ca urmare, distribuția valorilor în probe este aproximativ normală.

În urma parcurgerii acestei etape se poate spune că se îndeplinesc toate condițiile de aplicare ale testului t pentru o probă.

Ipotezele testului vor corespunde unei variante unilaterale, așa cum am stabilit în prima etapă:

$H_0: \mu \leq 100$
$H_1: \mu > 100$.

Cu ajutorul mediei ($\bar{x} = 105{,}87$), valorii de referință ($\mu_0 = 100$) și erorii standard a mediei ($s_{\bar{x}} = 2{,}142$) se calculează statistica testului:

$$t = \frac{105{,}87 - 100}{21{,}42} = 2{,}739 .$$

În continuare se verifică condiția testului. Pentru aceasta trebuie aflată valoarea critică a testului $t_{(\alpha, n-1)}$ din tabelul valorilor critice ale distribuției t (Student) (anexa 2). În cazul nostru, numărul gradelor de libertate este $n - 1 = 30 - 1 = 29$, iar nivelul de confidență $\alpha = 0{,}05$. Având în vedere că facem un test unilateral, $t_{(0{,}05,29)}$ va fi egală cu $t_{(0{,}1,29)}$, care s-ar folosi în cazul unui test unilateral, adică 1,699.

$2{,}739 > 1{,}699 \Rightarrow H_0$ se respinge, H_1 se acceptă
$p = 0{,}95$ sau 95%.

O altă modalitate de rezolvare a problemei constă în calcularea exactă a probabilității asociate valorii statisticii testului, adică să aflăm

proporția din distribuția Student exclusă de statistica testului t (anexa 3). Această proporție reprezintă probabilitatea (p) ca H_0 să fie adevărată, ceea ce înseamnă că probabilitatea ca H_0 să fie falsă sau ca H_1 să fie adevărată va fi $1-p$:

$p = 0{,}0052$, adică statistica testului $t = t_{(0{,}0052,29)}$.

Deci probabilitatea exactă ca media concentrației biomarker-ului în populația de pești să fie mai mică sau egală ca 100 (H_1) este de 0,0052 (0,52%), ceea ce înseamnă că probabilitatea să fie mai mare ca 100 este $1-0{,}0052 = 0{,}9948$ sau 99,48%.

Concluzia testului este cea surprinsă de H_1: media concentrației biomarker-ului în populația de pești din care a fost extrasă proba este semnificativ mai mare decât valoarea de referință.

Pentru a răspunde la întrebarea problemei, putem spune că apa din care a fost extrasă proba de 30 de pești este poluată.

7. TESTAREA DIFERENȚEI DINTRE DOUĂ PROBE

O astfel de inferență statistică se referă la compararea tendințelor centrale a două probe. Cele două probe pot fi prelevate din două populații diferite, caz în care se numesc **probe independente**. Denumirea este argumentată de faptul că eșantionarea valorilor primei probe nu influențează probabilitatea de extragere a valorilor celei de a doua probe, deoarece prelevarea se face din populații distincte.

Uneori, testarea diferenței se poate face pornindu-se de la **probe neindependente**. Într-un astfel de caz, cele două probe pot fi prelevate din aceeași populație sau se obțin prin investigarea unităților de probă de două ori: înainte și după aplicarea unui anumit tratament unităților de probă.

7.1. COMPARAREA A DOUĂ PROBE INDEPENDENTE

Când se face o astfel de comparație, se vizează de fapt compararea mediilor populațiilor din care au fost extrase probele. Deci mediile probelor reprezintă estimatori ai mediilor populațiilor din care au fost prelevate.

Cele mai utilizate teste ce pot fi folosite pentru compararea a două probe sunt **testul *t* (Student) pentru probe independente** și **testul *U* (Mann-Whitney)**.

7.1.1. Testul *t* (Student) pentru probe independente

Acesta este unul dintre cele mai utilizate teste parametrice ce se folosesc frecvent pentru o astfel de comparație în ecologie. Fiind un test parametric, prezintă o serie de condiții de aplicare:

1. cele două probe trebuie să fie prelevate aleator din două populații distincte;
2. variabila trebuie să fie apreciată pe o scală de interval sau de raport;
3. valorile din cele două probe trebuie să fie aproximativ normal distribuite.

În funcție de tipul de comparație, testul poate fi aplicat în variantă bilaterală sau unilaterală, în funcție de care se scriu ipotezele testului. Fie populația A din care se extrage aleator proba A și populația B din care se prelevează aleator proba B. Atunci ipotezele testului pot fi:

Bilateral	Unilateral	
$H_0: \mu_A = \mu_B$	$H_0: \mu_A \leq \mu_B$	$H_0: \mu_A \geq \mu_B$
$H_1: \mu_A \neq \mu_B$	$H_1: \mu_A > \mu_B$	$H_1: \mu_A < \mu_B$.

Statistica testului este următoarea:

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$\bar{x}_A$ – media valorilor din proba A
$\bar{x}_B$ – media valorilor din proba B
μ_A – media populației A μ_B – media populației B
s_A^2 – varianța probei A s_B^2 – varianța probei B
n_A – dimensiunea probei A n_B – dimensiunea probei B.

Atenție, diferența dintre mediile probelor μ_A și μ_B este 0 conform ipotezelor nule!

Valoarea critică cu care se compară statistica testului se află (anexa 2) în funcție de α și de numărul gradelor de libertate gl care se estimează conform ecuației Welch-Satterthwaite:

$$gl = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{1}{n_A - 1} \cdot \left(\frac{s_A^2}{n_A}\right)^2 + \frac{1}{n_B - 1} \cdot \left(\frac{s_B^2}{n_B}\right)^2}.$$

Există și o variantă mai simplă de estimare a gradelor de libertate:

$$gl = min(n_A, n_B) - 1 .$$

Condiția testului este:

dacă $|t| \geq t_{(\alpha,gl)} => H_0$ se respinge, H_1 se acceptă, $p = 1-\alpha$.

Probabilitatea asociată lui t sau H_0 se poate calcula exact (anexa 3), iar probabilitatea H_1 se află prin scăderea acesteia din 1:

$$p(H_1) = 1-p(H_0)\,.$$

Exemplul 7.1. La o probă extrasă aleator dintr-o populație de viperă de stepă (*Vipera ursinii moldavica*) s-a determinat numărul de perechi de plăci subcaudale. Este numărul de plăci subcaudale semnificativ mai mare la masculi decât la femele?

Femele (*f*):	31	32	29	29	29	25	29
	28	28	27	31	30	28	
Masculi (*m*):	35	37	38	38	38	36	38
	39	35	36	37	31	36	38
	36	37	37	38	37	32	

Dacă diferența dintre masculi și femele este semnificativă, înseamnă că este ca și cum ar proveni din două populații statistice diferite.

Datele îndeplinesc condițiile de aplicare ale testului t: datele au fost extrase aleator din populație, variabila este exprimată pe o scală de raport, datele sunt aproximativ normal distribuite ($\bar{x}_m = 36{,}45, Me_m = 37; \bar{x}_f = 28{,}923, Me_f = 29$). Intervalul $\bar{x} \pm s$ cuprinde aproximativ 62% din valorile femelelor și 85% din valorile masculilor. Pentru ca datele să fie mai normal distribuite, se realizează o normalizare a acestora prin logaritmare cu ajutorul logaritmului natural ($x' = ln(x)$).

f: 3,4340 3,4657 3,3673 3,3673 3,3673 3,2189 3,3673 3,3322 3,3322 3,2958 3,4340 3,4012 3,3322

m: 3,5553 3,6109 3,6376 3,6376 3,6376 3,5835 3,6376 3,6636 3,5553 3,5835 3,6109 3,4340 3,5835 3,6376 3,5835 3,6109 3,6109 3,6376 3,6109 3,4657

După transformare, intervalul $\bar{x} \pm s$ cuprinde aproximativ 85% din valorile femelelor și 90% din valorile masculilor

În continuare trebuie emise ipotezele testului. Cum întrebarea la care trebuie să răspundă testul este dacă numărul subcaudalelor la masculi este mai mare decât cel de la femele, atunci ipoteza alternativă va exprima această inegalitate.

$$H_0: \mu_m \leq \mu_f$$
$$H_1: \mu_m > \mu_f$$

Se calculează statistica testului:

$$t = \frac{3{,}3627 - 3{,}5944}{\sqrt{\frac{0{,}0042}{13} + \frac{0{,}0033}{20}}} = -10{,}482 .$$

Numărul gradelor de libertate se poate afla în două moduri:

$$gl = \frac{\left(\frac{0{,}0042}{13} + \frac{0{,}0033}{20}\right)^2}{\frac{1}{13-1}\left(\frac{0{,}0042}{13}\right)^2 + \frac{1}{20-1}\left(\frac{0{,}0033}{20}\right)^2} = 23{,}6 \approx 23$$
$$gl = min(13{,}20) - 1 = 13 - 1 = 12 .$$

Valoarea critică pentru un test unilateral, pentru 0,05 nivel de încredere și gl grade de libertate, se poate căuta în tabel (anexa 2) sau calcula (anexa 3):

$$t_{(0{,}05,23)} = 1{,}714 \qquad t_{(0{,}05,12)} = 1{,}782 .$$

Valoarea absolută a statisticii testului $|-10{,}482|$ este mai mare decât ambele valori critice, deci se respinge H_0 și se acceptă H_1, adică media numărului de subcaudale de la masculi este semnificativ mai mare decât cel de la femele, cu o probabilitate de 0,95 sau în 95% din cazuri.

Probabilitatea ca H_0 să fie adevărată (anexa 3) este de $1{,}6 \cdot 10^{-10}$ pentru 23 grade de libertate și $1{,}1 \cdot 10^{-7}$pentru 12 grade de libertate. În ambele situații, probabilitatea este foarte mică și se poate accepta ipoteza nulă a cărei probabilitate este de $1 - p(H_0)$, adică foarte mare.

7.1.2. Testul *U* (Mann-Whitney)

Acest test se utilizează ca alternativă neparametrică a testului *t* (Student) pentru probe independente. Unicele condiții ale testului *U* (Mann-Whitney) sunt:

1. probele trebuie să fie prelevate aleator din două populații distincte;
2. variabila trebuie să fie apreciată pe o scală ordinală, de interval sau de raport.

Ipotezele testului sunt în esență similare cu cele ale testului *t* pentru probe independente:

Bilateral	Unilateral	
$H_0: A = B$	$H_0: A \leq B$	$H_0: A \geq B$
$H_1: A \neq B$	$H_1: A > B$	$H_1: A < B$.

Aplicarea testului presupune ca valorile celor două probe să primească ranguri împreună. Pentru aceasta, valorile din ambele probe se ordonează crescător într-o singură serie. Valoarea cea mai mică va primi rangul 1, următoarea rangul 2 și așa mai departe. Valorile egale vor primi media rangurilor pe care le-ar fi primit dacă ar fi fost diferite (secțiunea 2.1, tab. 2.3).

Ulterior se însumează rangurile corespunzătoare valorilor fiecărei probe, obținându-se $\sum R_A$ – suma rangurilor corespunzătoare valorilor din proba *A* și, respectiv, $\sum R_B$ – suma rangurilor corespunzătoare valorilor din proba *B*. Cu valorile celor două sume și dimensiunile probelor (n_A și n_B) se calculează statisticile testului U_A și U_B:

$$U_A = n_A \cdot n_B + \frac{n_A(n_A+1)}{2} - \sum R_A$$
$$U_B = n_A \cdot n_B + \frac{n_B(n_B+1)}{2} - \sum R_B .$$

Suma celor două statistici ale testului trebuie să fie egală cu suma tuturor rangurilor, adică $U_A + U_B = n_A \cdot n_B$. De unde rezultă că, pentru simplificare, putem scrie una dintre statistici, să zicem U_B, în funcție de cealaltă:

$U_B = n_A \cdot n_B - U_A$.

Decizia privind respingerea sau acceptarea H_0 poate fi luată în două moduri, în funcție de dimensiunile celor două probe (n_A și n_B).

1. Dacă $n_A \leq 20$ și $n_B \leq 20$, atunci condiția testului este:
Dacă valoarea mai mică dintre statisticile testului este mai mică decât valoarea critică tabelată (anexa 2), atunci se respinge H_0 și se acceptă H_1.

Dacă $min(U_A, U_B) \leq U_{(\alpha,n_A,n_B)} \Rightarrow H_0$ se respinge, H_1 se acceptă pentru $p = 1-\alpha$.

Valoarea critică trebuie aleasă în funcție de varianta bilaterală sau unilaterală în care se aplică testul.

2. Dacă $n_A > 20$ și $n_B > 20$, atunci distribuția probabilistică a lui U este aproximativ normală și se poate face conversia la distribuția normală standard:
Pentru aceasta este nevoie să se calculeze media ($\bar{x}_U$) și deviația standard (s_U) ale lui U, dacă H_0 este adevărată:

$$\bar{x}_U = \frac{n_A \cdot n_B}{2}$$

$$s_U = \sqrt{\frac{n_A \cdot n_B (n_A + n_B + 1)}{12}} .$$

Cu acestea se calculează valoarea z corespunzătoare $N(0,1)$, unde U este una dintre cele două statistici ale testului (U_A sau U_B):

$$z = \frac{U - \bar{x}_U}{s_U} .$$

Dacă $|z| \geq z_{(0,95)} = 1,96 \Rightarrow H_0$ se respinge și se acceptă H_1 pentru p=0,95.

Exemplul 7.2. Să se răspundă la întrebarea de la exemplul 7.1, considerându-se că una dintre sau ambele condiții 2 și 3 ale testului Student pentru probe independente nu sunt îndeplinite.

În acest caz se utilizează alternativa neparametrică a testului Student, adică testul Mann-Whitney. Pentru a realiza acest test se dau ranguri valorilor din cele două probe împreună, după metoda prezentată în secțiunea 2.1.

Sex	*f*	*f*	*f*	*F*	*f*	*f*	*f*	*f*	*f*	*f*	*f*	*f*	*f*	*m*	*m*	*m*	*m*
x	25	27	28	28	28	29	29	29	29	30	31	31	32	31	32	35	35
R_i	1	2	3	4	5	6	7	8	9	10	11	12	14	13	15	16	17
R_x	1	2	4	4	4	7,5	7,5	7,5	7,5	10	12	12	14,5	12	14,5	16,5	16,5

continuare

m	*m*	*m*	*M*	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*
36	36	36	36	37	37	37	37	37	38	38	38	38	38	38	39
18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
19,5	19,5	19,5	19,5	24	24	24	24	24	29,5	29,5	29,5	29,5	29,5	29,5	33

Ipotezele testului sunt similare cu cele de la exemplul 7.1:

$H_0: m \leq f$
$H_1: m > f$.

Se calculează suma rangurilor valorilor de la femele:

$$\sum R_{x_f} = 1 + 2 + 4 + \cdots + 14{,}5 = 93{,}5 .$$

Statisticile testului vor fi:

$$U_f = 13 \cdot 20 + \frac{13(13+1)}{2} - 93{,}5 = 257{,}5$$
$$U_m = 13 \cdot 20 - 257{,}5 = 2{,}5 .$$

Deoarece dimensiunea probelor este mai mică sau egală cu 20, trebuie aflată valoarea critică (anexa 2). Cum testul se aplică în variantă unilaterală, atunci valoarea critică din tabelul cu valorile critice $U_{(\alpha,n_A,n_B)}$ din anexa 2 va fi considerată pentru un nivel de confidență de 0,025:

$$U_{(0,025,13,20)} = 76 .$$

De fapt, pentru testul Mann-Whitney există un interval de acceptare a H_0. Suma limitelor acestui interval este egală cu produsul dintre dimensiunile probelor, în cazul exemplului fiind egală cu $13 \cdot 20 = 260$. Cum una dintre limite este valoarea critică egală cu 76, înseamnă că cealaltă este egală cu $260 - 76 = 184$. Deci intervalul de acceptare a H_0 are limita inferioară 76 și limita superioară 184. Cele două statistici calculate sunt în afara acestui interval, deci ipoteza nulă se respinge și se acceptă în consecință ipoteza alternativă.

U_m	LI	LS	U_f
2,5	76	184	257,5
$H_1 \rightarrow$	$\leftarrow H_0 \rightarrow$		$\leftarrow H_1$

Concluzia este că numărul de subcaudale la masculi este semnificativ mai mare decât la femele, cu o probabilitate de 0,975 sau în 97,25% din cazuri.

7.2. COMPARAREA A DOUĂ PROBE NEINDEPENDENTE

Aceste comparații au drept scop evidențierea efectului unui anumit tratament asupra valorilor variabilei investigate. Tratamentul poate fi acțiunea unei substanțe, a unui factor de mediu etc. ce ar putea modifica valorile individuale ale unei variabile.

În cazul unor astfel de comparații, probele trebuie să fie neindependente, deoarece se presupune că astfel s-ar elimina posibilitatea apariției unor diferențe semnificative datorate deosebirilor dintre populații. Astfel, diferențele, dacă apar, vor reprezenta efectul tratamentului asupra unităților de probă.

Atunci când este posibil, este bine ca cele două probe să rezulte în urma a două investigații repetate asupra unităților de probă: prima, înainte de aplicarea tratamentului, și a doua, după aplicarea acestuia. Astfel, rezultă câte o pereche de valori pentru fiecare unitate de probă.

Existența perechilor de valori este absolut necesară în cazul testelor ce compară probe neindependente, motiv pentru care acestea se mai numesc și teste pentru perechi de valori sau pentru observații perechi.

Testele cele mai uzuale care se folosesc pentru astfel de observații sunt **testul *t* (Student) pentru perechi de observații** și **testul *T* (Wilcoxon).**

7.2.1. Testul *t* (Student) pentru perechi de observații

Ca și testul *t* pentru probe independente, și acesta este un test parametric care necesită ca datele să îndeplinească o serie de condiții de aplicare:

1. probele trebuie să fie extrase aleator din aceeași populație sau să provină în urma unor investigații repetate asupra acelorași unități de probă, iar dimensiunea probelor trebuie să fie aceeași astfel încât să existe perechi de observații;
2. variabila trebuie să fie apreciată pe o scală de interval sau de raport;
3. distribuția valorilor în probe trebuie să fie aproximativ normală.

În funcție de întrebarea la care testul trebuie să răspundă, se scriu ipotezele în variantă bilaterală sau unilaterală. În esență, ipotezele se referă la media diferențelor (μ_D) dintre perechile de valori în populație:

Bilateral	Unilateral	
$H_0: \mu_D = 0$	$H_0: \mu_D \leq 0$	$H_0: \mu_D \geq 0$
$H_1: \mu_D \neq 0$	$H_1: \mu_D > 0$	$H_1: \mu_D < 0$.

Dacă se notează diferența dintre valorile unei perechi, iar fiecare dintre valorile perechii aparțin probei A și, respectiv, probei B, atunci:

$$D = x_A - x_B \ .$$

Dacă media acestor diferențe în populație este zero, atunci înseamnă că tratamentul aplicat nu a modificat semnificativ valorile variabilei. O medie a diferențelor mai mare ca zero arată că tratamentul a dus la creșterea semnificativă a valorilor, în timp ce una mai mică decât zero arată că tratamentul a scăzut semnificativ valorile.

Principalul estimator al mediei populaționale a diferențelor este media diferențelor dintre perechile de valori din probe ($\overline{D}$). Deci se calculează diferențele corespunzătoare fiecărei perechi de valori, iar suma diferențelor se împarte la numărul perechilor de valori (n) sau la numărul valorilor dintr-o probă.

$$\overline{D} = \frac{\sum D}{n}$$

O altă valoare necesară pentru aflarea statisticii testului este eroarea standard a mediei diferențelor ($s_{\overline{D}}$), care se calculează împărțind deviația standard a mediei diferențelor (s_D) la radical din numărul perechilor de observații (n):

$$s_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{n}}{n-1}} \qquad s_{\overline{D}} = \frac{s_D}{\sqrt{n}}.$$

Statistica testului (t) va fi egală cu raportul dintre media diferențelor ($\overline{D}$) și eroarea standard a mediei diferențelor ($s_{\overline{D}}$):

$$t = \frac{\overline{D}}{s_{\overline{D}}}.$$

Condiția testului, atât pentru varianta bilaterală, cât și pentru variantele unilaterale, este:

dacă $|t| \geq t_{(\alpha,n-1)}$ => H_0 se respinge, H_1 se acceptă, pentru p=1–α.

Valoarea critică se poate afla din tabele (anexa 2) sau se poate calcula (anexa 3).

Când se calculează probabilitatea asociată valorii t (anexa 3) (adică probabilitatea ca ipoteza nulă să fie adevărată), atunci ipoteza nulă se respinge dacă $p(H_0)$ este mai mic de nivelul α și se va accepta ipoteza alternativă pentru o probabilitate de $1 - p(H_0)$.

Exemplul 7.3. Într-un experiment s-a urmărit reacția gândacilor de bucătărie la lumină. Pentru aceasta s-au pus 12 indivizi selectați aleatoriu în 12 incinte cu pereți transparenți și la aprinderea luminii s-a urmărit câte secunde dintr-un minut a petrecut fiecare gândac în mijlocul incintei și lângă pereții incintei. Au gândacii o tendință semnificativă de a se ascunde în condiții de lumină?

scunde în mijloc (*A*)	28	17	33	24	15	23	21	21	21	32	27	34
secunde lângă perete (*B*)	32	43	27	36	45	37	39	39	39	28	33	26

În acest exemplu, sunt prezentate două probe rezultate prin investigarea acelorași unități de probă (cei 12 gândaci) de două ori: pentru fiecare gândac există timpul petrecut în mijlocul incintei și timpul petrecut lângă pereți. Deci probele nu sunt independente și pentru fiecare unitate de probă (gândac) există o pereche de valori (timpul petrecut în mijlocul incintei și timpul petrecut lângă pereții acesteia).

În continuare trebuie verificate condițiile de aplicare ale testului Student pentru perechi de valori: probele sunt aleatoare și există observații perechi, variabila este exprimată pe o scală de raport, probele au o distribuție aproximativ normală ($\bar{x}_A = 24{,}67, Me_A = 23{,}5; \bar{x}_B = 35{,}33, Me = 36{,}5$). Condițiile sunt îndeplinite.

Pentru a emite ipotezele testului, trebuie analizată întrebarea problemei: dacă gândacii ar avea o tendință semnificativă să se ascundă, atunci ar trebui ca timpii petrecuți lângă pereți să fie semnificativ mai mari decât cei petrecuți în centru. Dacă se calculează diferența $D = x_B - x_A$, atunci diferențele pozitive trebuie să fie dominante, fapt ce trebuie surprins de ipoteza alternativă.

$H_0: \mu_D \leq 0$
$H_1: \mu_D > 0$

Se calculează diferențele dintre valorile fiecărei perechi:

x_A	28	17	33	24	15	23	21	21	21	32	27	34
x_B	32	43	27	36	45	37	39	39	39	28	33	26
$D = x_B - x_A$	4	26	–6	12	30	14	18	18	18	–4	6	–8

Se calculează media diferențelor și eroarea standard a mediei diferențelor care sunt necesare pentru aflarea statisticii testului:

$$\overline{D} = \frac{128}{12} = 10{,}67$$

$$s_D = \sqrt{\frac{3056 - \frac{16384}{12}}{12-1}} = 12{,}397$$

$$s_{\overline{D}} = \frac{12{,}397}{\sqrt{12}} = 3{,}579$$

$$t = \frac{10{,}67}{3{,}579} = 2{,}98\,.$$

Se află valoarea critică pentru testul unilateral, pentru un nivel de confidență de 0,05 și $12 - 1$ grade de libertate (anexa 2 sau anexa 3):

$$t_{(0{,}05;11)} = 1{,}796\,.$$

Valoarea statisticii testului este mai mare decât valoarea critică, deci se respinge ipoteza nulă și se acceptă ipoteza alternativă, adică numărul de secunde petrecut lângă pereții incintei este semnificativ mai mare decât numărul de secunde petrecute în centrul incintei, cu o probabilitate de 0,95 sau în 95% din cazuri.

Probabilitatea ca ipoteza nulă să fie adevărată (anexa 3) este de 0,0063, deci probabilitatea ca ipoteza nulă să nu fie adevărată și să fie adevărată ipoteza alternativă este de $1 - 0{,}0063 = 0{,}9937$.

7.2.2. Testul *T* (Wilcoxon)

Acest test reprezintă alternativa neparametrică a testului *t* pentru observații perechi, deci se utilizează pentru două probe neindependente sau pentru observații perechi realizate asupra unităților de probă, înainte sau după aplicarea unui tratament. Se folosește atunci când nu se respectă

condiția 2 sau 3 a testului Student pentru perechi de observații. Condițiile de aplicare a testului Wilcoxon sunt:

1. probele trebuie să fie extrase aleator din aceeași populație sau să provină în urma unor investigații repetate asupra aceloraşi unități de probă, iar dimensiunea probelor trebuie să fie aceeași astfel încât să existe perechi de observații;
2. variabila trebuie să fie apreciată pe o scală ordinală, de interval sau de raport.

În cazul acestui test, se calculează diferența (D) dintre observațiile fiecărei perechi de valori din cele două probe.

$$D = x_B - x_A$$

Diferențele în modul primesc apoi ranguri ($|D| \rightarrow R_D$) la fel ca în cazul testului U Mann-Whitney. În dreptul fiecărui rang se specifică între paranteze semnul diferenței corespunzătoare.

Valorile ale căror diferențe sunt nule se elimină din analiză, ceea ce atrage după sine reducerea corespunzătoare a gradelor de libertate.

Ulterior se calculează suma rangurilor diferențelor pozitive ($\sum R_{D>0}$) și suma rangurilor diferențelor negative ($\sum R_{D<0}$).

Ipotezele care se scriu în acest test sunt în esență similare cu cele ale testului t pentru probe neindependente, cu excepția faptului că se referă la sumele rangurilor diferențelor:

Bilateral	Unilateral	
$H_0: \sum R_{D>0} = \sum R_{D<0}$	$H_0: \sum R_{D>0} \leq \sum R_{D<0}$	$H_0: \sum R_{D>0} \geq \sum R_{D<0}$
$H_1: \sum R_{D>0} \neq \sum R_{D<0}$	$H_1: \sum R_{D>0} > \sum R_{D<0}$	$H_1: \sum R_{D>0} < \sum R_{D<0}$.

Ce reprezintă aceste diferențe? Dacă suma rangurilor diferențelor pozitive este mai mare decât cea a rangurilor diferențelor negative, înseamnă că există mai multe diferențe pozitive, deci valorile din proba B, influențate de tratament, sunt în general mai mari decât cele din proba A. Aceasta înseamnă că tratamentul aplicat a dus la creșterea valorilor. Dacă suma rangurilor diferențelor pozitive este mai mică decât cea a rangurilor

diferențelor negative, înseamnă că sunt mai multe diferențe negative și valorile din proba A sunt în general mai mari decât cele din B. Deci tratamentul aplicat a determinat scăderea valorilor probei B. Dacă cele două sume sunt egale, înseamnă că există diferențe negative și pozitive în egală măsură și tratamentul nu a modificat semnificativ valorile din proba B.

Statisticile testului (T și T') sunt reprezentate de suma rangurilor diferențelor pozitive și suma rangurilor diferențelor negative:

$$T = min\,(\sum R_{D>0}\,,\sum R_{D<0})$$
$$T' = \frac{n(n+1)}{2} - T \qquad \text{sau} \qquad T' = max\,(\sum R_{D>0}\,,\sum R_{D<0})\,.$$

Inspectarea formulelor de mai sus arată că suma celor două statistici este egală cu suma rangurilor tuturor diferențelor (diferite de 0) atât pozitive, cât și negative:

$$T + T' = \frac{n(n+1)}{2}\,.$$

Decizia de acceptare sau de respingere a H_0 se ia în funcție de numărul gradelor de libertate (n).

1. Dacă $n \leq 33$ (sau decât 25), condiția testului constă în compararea statisticilor cu valoarea critică $T_{(\alpha,n)}$ care se găsește în tabele (anexa 2) și se alege în funcție de varianta în care se aplică testul și numărul gradelor de libertate egal cu numărul diferențelor nenule.

Dacă T sau $T' \leq T_{(\alpha,n)}$ $\Rightarrow$ H_0 se respinge, H_1 se acceptă, pentru $p = 1-\alpha$.

Dacă se acceptă H_1 în cazul unei variante bilaterale, concluzia este că tratamentul a modificat semnificativ variabila. În cazul variantei unilaterale dreapta ($H_1: \sum R_{D>0} > \sum R_{D<0}$), concluzia va fi că tratamentul a determinat o creștere a valorilor variabilei, iar în cazul variantei unilaterale stânga ($H_1: \sum R_{D>0} < \sum R_{D<0}$), că acesta le-a scăzut.

2. Dacă $n > 33$, atunci distribuția T poate fi aproximată prin cea normală standard. Pentru a putea calcula valoarea z trebuie mai întâi calculate media $\bar{x}_T$ și deviația standard s_T.

$$\bar{x}_T = \frac{n(n+1)}{4}$$

$$s_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$z = \frac{T - \bar{x}_T}{s_T}$$

Condiția testului este:

dacă $|z| \geq z_{(0,95)} = 1,96 \Rightarrow H_0$ se respinge, H_1 se acceptă, pentru $p = 0,95$

Exemplul 7.4. Să se rezolve problema din exemplul 7.4, considerându-se că una dintre ultimele două condiții ale testului Student pentru perechi de observații nu este îndeplinită.

În acest caz se aplică alternativa neparametrică a testului Student pentru perechi de observații, adică testul Wilcoxon. Pentru efectuarea acestui test se calculează diferențele dintre valorile fiecărei perechi. Modulelor diferențelor li se dau ranguri, iar în dreptul fiecărui rang se specifică semnul diferenței corespunzătoare.

Ipotezele testului sunt asemănătoare cu cele de la exemplul 7.3. Dacă diferențele dintre valorile fiecărei perechi se calculează scăzând din timpul petrecut lângă perete pe cel petrecut în mijlocul incintei, atunci, dacă există o tendință semnificativă de ascundere la gândacii de bucătărie, ar trebui să se obțină mai multe diferențe pozitive și suma rangurilor diferențelor pozitive ar trebui să fie mai mare decât cea a diferențelor negative. Acest aspect trebuie surprins în ipoteza alternativă.

$$H_0: \sum R_{D>0} \leq \sum R_{D<0}$$
$$H_1: \sum R_{D>0} > \sum R_{D<0}$$

Secunde în centru (A)	Secunde lângă perete (B)	$B-A$	$\lvert B-A\rvert$	R_i	R_D	semnul diferenței
28	32	4	4	1	1,5	+
32	28	–4	4	2	1,5	–
33	27	–6	6	3	3,5	–
27	33	6	6	4	3,5	+
34	26	–8	8	5	5	–
24	36	12	12	6	6	+
23	37	14	14	7	7	+
21	39	18	18	8	9	+
21	39	18	18	9	9	+
21	39	18	18	10	9	+
17	43	26	26	11	11	+
15	45	30	30	12	12	+

Se calculează statisticile testului:

$T = 1{,}5 + 3{,}5 + 5 = 10$

$T' = \frac{12(12+1)}{2} - 10 = 68$.

Deoarece numărul perechilor de observații este mai mic de 33, se află valoarea critică pentru varianta bilaterală în funcție de nivelul de confidență și de numărul gradelor de libertate:

$T_{(0,05,12)} = 17$.

Și pentru testul Wilcoxon există un interval de acceptare a H_0. Suma limitelor acestui interval este egală cu suma tuturor rangurilor:

$\sum R_D = \frac{12(12+1)}{2} = 78$.

Cum una dintre limite este reprezentată de valoarea critică egală cu 17, înseamnă că cealaltă este egală cu $78 - 17 = 61$. Deci intervalul de acceptare a H_0 are limita inferioară 17 și limita superioară 61. Cele două statistici calculate sunt în afara acestui interval, deci ipoteza nulă se respinge

și se acceptă în consecință ipoteza alternativă.

T	LI	LS	T'
10	17	61	68
$H_1 \rightarrow$	$\leftarrow H_0 \rightarrow$		$\leftarrow H_1$

Concluzia este că timpul petrecut lângă perete este semnificativ mai mare decât cel petrecut în centrul incintei, cu o probabilitate de 0,95 sau în 95% din cazuri.

8. TESTAREA DIFERENȚELOR DINTRE TREI SAU MAI MULTE PROBE

Există situații în care, în cadrul unor investigații ecologice, este necesară testarea semnificației diferențelor dintre trei sau mai multe probe din punctul de vedere al unei variabile. Dacă datele îndeplinesc condițiile de aplicare ale unor teste parametrice, atunci o astfel de situație s-ar putea rezolva printr-o serie de teste t care să verifice semnificația diferenței pentru fiecare pereche de probe până la epuizarea tuturor combinațiilor unice posibile. De exemplu, dacă există trei probe A, B și C, atunci se face câte un test Student pentru verificarea semnificației diferenței dintre A și B, dintre A și C și dintre B și C. Dacă însă se compară 10 probe în loc de 3, atunci analiza devine ceva mai dificilă, pentru că trebuie efectuate 45 de teste Student. Pe lângă aceasta, într-o astfel de situație mai apare si un aspect statistic în defavoarea unei asemenea abordări.

Așa cum s-a arătat în secțiunea 5.5, orice test presupune un risc α de a efectua o eroare de genul I, ce constă în respingerea unei ipoteze nule care în realitate este adevărată. Cum valoarea cea mai uzuală a lui α este 0,05, înseamnă că pentru cele 45 de teste Student riscul apariției unei astfel de erori crește la $45 \cdot 0{,}05 = 2{,}25$. Conform acestei valori, înseamnă că există șanse mari de a ajunge la cel puțin două concluzii greșite cu privire la diferențele dintre cele zece probe. O posibilă rezolvare ar consta în scăderea valorii α de la 0,05 la 0,01, ceea ce ar micșora riscul în cazul exemplului la $45 \cdot 0{,}01 = 0{,}45$. Însă scăzând valoarea nivelului de semnificație, crește riscul β de a comite o eroare de genul II, adică de a accepta o ipoteză nulă care în realitate să fie falsă.

Aceste neajunsuri care pot să apară când se compară trei sau mai multe probe pot fi depășite cu ajutorul **analizei variației**. Această tehnică de analiză statistică este simbolizată prin acronimul **ANOVA** provenit de la denumirea sa în limba engleză (**AN**alysis **O**f **VA**riance), acronim pe care îl vom folosi în continuare pentru a ne referi la această tehnică.

ANOVA este o tehnică versatilă ce poate fi utilizată pentru

compararea a trei sau mai multor probe, grupate după unul sau mai mulți factori, în cadrul unui singur test.

8.1. PRINCIPIUL ANOVA

În esență, ANOVA este o tehnică ce permite descompunerea variabilității unui set de probe în componentele sale. Dacă ne imaginăm că avem de comparat trei probe, atunci **variabilitatea totală** a acestora va fi dată de variabilitatea valorilor individuale față de media fiecărei probe, adică **variabilitatea internă** a probelor, și de variabilitatea dintre probe datorată diferențelor dintre mediile populațiilor din care au fost extrase probele, adică **variabilitatea externă**. Dacă descriptorul folosit al variabilității este varianța (s^2), atunci descompunerea variabilității poate fi rezumată prin relația:

$s_t^2 = s_{ext}^2 + s_{int}^2$
s_t^2 – varianța totală;
s_{ext}^2 – varianța externă sau dintre probe;
s_{int}^2 – varianța internă sau din probe.

În ANOVA varianțele nu se folosesc ca atare, ci se consideră ca fiind sume de pătrate medii ($\overline{SP}$) obținute prin împărțirea sumelor de pătrate (SP) la numărul gradelor de libertate (gl).

$$s^2 = \overline{SP} = \frac{SP}{gl}$$

Astfel vom vorbi de sume de pătrate medii, sume de pătrate și grade de libertate, totale, externe și interne.

Dacă probele au fost extrase aleatoriu din populații normal distribuite cu medii și varianțe egale, atunci varianța internă va fi aceeași cu cea externă. În cazul în care probele provin din populații cu medii și varianțe diferite, atunci diferențele dintre probe vor avea drept cauză principală variabilitatea sau varianța externă. Statistica ANOVA, notată cu F, reprezintă tocmai raportul dintre varianța externă și cea internă și are o distribuție particulară Snedecor-Fisher în funcție de nivelul de semnificație (α), de gradele de libertate externe (gl_{ext}) și de gradele de libertate interne

(gl_{int}). Cu cât valoarea lui F va fi mai mare, cu atât variabilitatea totală va rezulta mai mult din variabilitatea externă și mai puțin din cea internă, și invers.

ANOVA este o tehnică statistică parametrică, deci presupune ca datele analizate să îndeplinească următoarele condiții:

1. probele trebuie să fie prelevate aleator;
2. variabila trebuie să fie apreciată pe o scală de interval sau de raport;
3. datele din probe trebuie să fie aproximativ normal distribuite;
4. variațele interne trebuie să nu difere semnificativ.

Ipotezele generale care se pot scrie în ANOVA sunt:

H_0: probele au fost prelevate din populații normal distribuite cu variațe și medii egale.

H_1: deoarece se presupune că variațele populațiilor sunt egale, probele au fost prelevate din populații cu medii diferite.

Dacă datele nu respectă condițiile enumerate mai sus, se poate apela la transformări ale valorilor variabilelor sau la alternative neparametrice ale ANOVA.

Se observă că se face referire la egalitatea variațelor interne ale probelor atât la nivelul condițiilor de aplicare ale ANOVA, cât și la nivelul ipotezelor generale. Deci, după verificarea îndeplinirii condițiilor comune testelor parametrice, trebuie testat dacă variațele interne ale probelor diferă semnificativ.

8.1.1. Testarea omogenității variației interne

În acest sens se pot utiliza două teste diferite. Cel mai rapid și mai simplu este **testul F_{max}** sau **Hartley**. Acesta constă în calcularea raportului între varianța cea mai mare (s^2_{max}) și cea mai mică (s^2_{min}) dintre variațele probelor analizate.

$$F_{max} = \frac{s^2_{max}}{s^2_{min}}$$

Valoarea F_{max} se compară cu o valoare critică pentru un anumit α, pentru un anumit număr de probe (k) și grade de libertate ($n-1$) (anexa 2).

Dacă $F_{max} < F_{max_{(\alpha,k,n-1)}} \Rightarrow$ nu există diferențe semnificative între variațele interne ale probelor pentru $p = 1 - \alpha$.

Acest test se poate folosi doar dacă toate probele conțin același număr de valori ($n_1 = n_2 = \cdots = n_k$).

Dacă probele au dimensiuni diferite, atunci se recomandă aplicarea testului **Bartlett**. Statistica testului Bartlett este una de tip χ^2 și se calculează pornind de la gradele de libertate ale fiecărei probe ($n_i - 1$, unde i ia valori de la 1 la numărul de probe k), de la varianța fiecărei probe (s_i^2) și de la varianța medie balansată (s_w^2).

$$s_w^2 = \frac{\sum s_i^2(n_i-1)}{\sum(n_i-1)}$$

$$\chi^2 = ln(s_w^2) \cdot \sum(n_i - 1) - \sum[ln(s_i^2)(n_i - 1)]$$

Statistica testului se compară cu o valoare critică χ^2 în funcție de un anumit nivel de confidență α și numărul de probe minus unu ($k-1$) grade de libertate (anexele 2 și 3).

Dacă $\chi^2 < \chi^2_{(\alpha,k-1)} \Rightarrow$ nu există diferențe semnificative între variațele interne ale probelor pentru $p = 1 - \alpha$.

Probabilitatea asociată valorii statisticii testului poate fi calculată și exact (anexa 3). Când se calculează probabilitatea asociată valorii χ^2 se acceptă că nu există diferențe semnificative între variațele probelor dacă p este mai mare de nivelul α.

Exemplul 8.1. S-a urmărit numărul de exemplare ale unei plante în câte 12 suprafețe de probă din 4 zone diferite (notate de la A la D). Urmează să se realizeze o analiză a variației și este nevoie să se testeze dacă variațele probelor diferă semnificativ. Datele sunt următoarele:

A	*B*	*C*	*D*
28	32	54	59
41	10	46	63
25	12	39	71
13	25	21	53
25	38	41	66
33	14	25	74
13	37	23	35
30	43	46	62
21	16	29	53
22	20	52	49
12	28	31	53
29	33	57	29

Se examinează normalitatea distribuției datelor în probe.

$$\bar{x}_A = 24{,}33 \quad Me_A = 25$$
$$\bar{x}_B = 25{,}66 \quad Me_B = 26{,}5$$
$$\bar{x}_C = 38{,}66 \quad Me_C = 40$$
$$\bar{x}_D = 55{,}58 \quad Me_D = 56$$

Se poate considera că probele au o distribuție aproximativ normală.

În continuare se verifică omogenitatea varianței cu ajutorul testului F_{max}, pentru efectuarea căruia este necesară calcularea variațelor probelor:

$$s_A^2 = \frac{7952 - \frac{(292)^2}{12}}{12-1} = 76{,}97 \qquad s_B^2 = \frac{9280 - \frac{(308)^2}{12}}{12-1} = 124{,}97$$

$$s_C^2 = \frac{19700 - \frac{(464)^2}{12}}{12-1} = 159{,}88 \qquad s_D^2 = \frac{39061 - \frac{(667)^2}{12}}{12-1} = 180{,}63\,.$$

Se calculează statistica testului:

$$F_{max} = \frac{180{,}63}{76{,}97} = 2{,}347\,.$$

Valoarea critică a testului se află în funcție de nivelul de confidență, numărul probelor și numărul de valori per probă minus unu (anexa 2). Pentru că valoarea pentru 4 și 11 grade de libertate lipsește din tabel, vom considera valoarea pentru 4 și 10:

$$F_{max_{(0{,}05,4,10)}} = 5{,}67\,.$$

Valoarea calculată este mai mică decât valoarea critică, deci se poate considera că variațele probelor nu diferă semnificativ sau varianța internă este omogenă.

Dacă se folosește testul Bartlett pentru aceleași date, trebuie calculată media balansată a varianței:

$$s_w^2 = \frac{76{,}97(12-1)+124{,}97(12-1)+159{,}88(12-1)+180{,}63(12-1)}{(12-1)+(12-1)+(12-1)+(12-1)} = 135{,}612\,.$$

Se calculează suma produselor dintre varianța logaritmată a probei și numărul gradelor de libertate:

$$\sum[ln(s_i^2)(n_i - 1)] = ln(76{,}97)(12-1) + ln(124{,}97)(12-1) + ln(159{,}88)(12-1) + ln(180{,}63)(12-1) = 213{,}866\,.$$

Se calculează statistica testului:

$$\chi^2 = ln(135{,}612) \cdot 44 - 213{,}866 = 2{,}165\,.$$

Se află valoarea critică (anexa 2 sau anexa 3):

$$\chi^2_{(0{,}05,4-1)} = 7{,}815\,.$$

Statistica testului este mai mică decât valoarea critică, deci varianța internă poate fi considerată omogenă. Probabilitatea ca să nu existe diferențe semnificative între variațele probelor este de 0,54, adică mai mare de 0,05.

8.2. TIPURI DE ANOVA

ANOVA este una dintre cele mai versatile tehnici de statistică inferențială, putând fi aplicată în numeroase variante în funcție de planificarea investigațiilor, atât în teren, cât și în laborator.

În cadrul ANOVA, gruparea valorilor variabilei investigate se poate face după unul sau mai mulți **factori**. Un factor reprezintă un grup de **tratamente** similare, care la rândul lor reprezintă **niveluri ale factorului** respectiv. Această terminologie își are originea în experimentele din agricultură, însă în prezent a căpătat o extindere mult mai mare, depășind semnificația inițială. Astfel, într-un experiment, nivelurile unui factor sau tratamentele pot fi obținute în urma unor manipulări. De exemplu, trei loturi de animale primesc fiecare un anumit tip de hrană și se urmărește efectul

acestuia asupra greutății corporale. În acest exemplu, factorul este reprezentat de hrană, iar tratamentele sau nivelurile factorului sunt tipurile de hrană. Datele vor fi reprezentate de greutățile animalelor și vor fi grupate în probe, în funcție de tipul de hrană. Dacă însă se urmărește o anumită variabilă în trei populații conspecifice din trei locații diferite, atunci factorul va fi locația în general, iar nivelurile acestuia vor fi reprezentate de locațiile fiecărei populații în parte. Valorile variabilei urmărite vor fi grupate în trei probe, în funcție de locația fiecărei populații. În concluzie, factorul poate fi orice factor ecologic ale cărui niveluri („tratamente") diferă de la o populație la alta.

În funcție de numărul factorilor după care se grupează datele, ANOVA poate fi unifactorială dacă ia în considerație un singur factor, sau multifactorială, dacă se iau în calcul mai mulți factori. Modelele cele mai frecvent utilizate în cercetările ecologice sunt modelul **unifactorial** și cel **bifactorial**.

Un alt aspect caracteristic pentru ANOVA bifactorială este **interacțiunea factorilor**. Dacă există interacțiune între factori, atunci se utilizează **modelul bifactorial cu replicare** (cu număr egal de observații în celulă), iar dacă nu există interacțiune între factori atunci se folosește **modelul bifactorial fără replicare** (cu o singură observație în celulă sau cu observații repetate). Din perspectiva ANOVA, interacțiunea reprezintă o parte a variabilității totale, datorată modificărilor unui factor, și este legată de variabilitatea unui alt factor sau a unei combinații de alți factori.

8.2.1. ANOVA unifactorială

În cazul acestui model datele sunt grupate în probe în funcție de un singur factor (F). Astfel, probele coincid nivelurilor factorului sau tratamentelor ($F_i = F_1 \ldots F_k$) (tab. 8.1).

Conform acestui model orice observație poate fi definită ca:

$$\text{Obs.} = \begin{matrix}\text{Media}\\ \text{generală}\end{matrix} + \text{Efectul } F_i + \begin{matrix}\text{Eroarea}\\ \text{întâmplătoare.}\end{matrix}$$

Efectul nivelurilor factorului dă variabilitatea externă (dintre probe), iar eroarea întâmplătoare este rezultatul variabilității interne (din cadrul probelor).

$$s_t^2 = s_{ext}^2 + s_{int}^2$$
$$s^2 = \overline{SP} = \frac{SP}{gl}$$

Din aceste două relații rezultă:

$$SP_t = SP_{ext} + SP_{int}$$
$$gl_t = gl_{ext} + gl_{int}\ .$$

Tabelul 8.1. Distribuția valorilor în probe în ANOVA unifactorială

F			
F_1	F_2	...	F_k
x_{1_1}	x_{2_1}		x_{k_1}
x_{1_2}	x_{2_2}		x_{k_2}
⋮	⋮	...	⋮
$x_{1_{n_1}}$	$x_{2_{n_2}}$		$x_{k_{n_k}}$

Ipotezele care se testează prin intermediul acestui model pot fi formulate diferit, în funcție de sensul care se atribuie noțiunii de tratament. În cazul unor date obținute în urma unui experiment, ipotezele sunt de forma:

H_0: nu există diferențe semnificative între efectele tratamentelor asupra variabilei;
H_1: diferențele dintre efectele tratamentelor asupra variabilei sunt semnificative.

În cazul unor date obținute în urma realizării unor observații efectuate asupra unei variabile în populații diferite, ipotezele pot fi formulate astfel:

H_0: mediile populațiilor din care s-au extras probele nu diferă semnificativ;
H_1: mediile populațiilor din care s-au extras probele diferă semnificativ.

Dacă se aplică ANOVA unifactorială pentru k probe, trebuie parcurse următoarele etape:

1. Se calculează suma de pătrate a tuturor valorilor ($\sum x_t^2$) prin adunarea sumelor de pătrate a valorilor pentru fiecare probă ($\sum x_i^2$):

$$\sum x_t^2 = \sum x_1^2 + \sum x_2^2 + \cdots + \sum x_k^2 .$$

2. Se calculează pătratul sumei totale ($(\sum x_t)^2$) prin adunarea sumelor valorilor din fiecare probă ($\sum x_i$), urmată de ridicarea la pătrat:

$$(\sum x_t)^2 = (\sum x_1 + \sum x_2 + \cdots + \sum x_k)^2 .$$

3. Se calculează numărul total de valori din toate probele (n_t) prin însumarea dimensiunilor tuturor probelor (n_i):

$$n_t = n_1 + n_2 + \cdots + n_k .$$

4. Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărei probe:

$$\sum \left[\frac{(\sum x_i)^2}{n_i}\right] = \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \cdots + \frac{(\sum x_k)^2}{n_k} .$$

5. Se calculează sumele de pătrate totală (SP_t), externă (SP_{ext}) și internă (SP_{int}):

$$SP_t = \sum x_t^2 - \frac{(\sum x_t)^2}{n_t} \qquad SP_{ext} = \sum \left[\frac{(\sum x_i)^2}{n_i}\right] - \frac{(\sum x_t)^2}{n_t}$$
$$SP_{int} = SP_t - SP_{ext} .$$

6. Se calculează numărul gradelor de libertate totale (gl_t), externe (gl_{ext}) și interne (gl_{int}):

$$gl_t = n_t - 1 \qquad gl_{ext} = k - 1$$
$$gl_{int} = gl_t - gl_{ext} = n_t - k .$$

7. Sumele de pătrate medii ($\overline{SP}$) se calculează împărțind sumele de pătrate (SP) la gradele de libertate corespunzătoare (gl):

$$\overline{SP}_{ext} = \frac{SP_{ext}}{gl_{ext}} \qquad \overline{SP}_{int} = \frac{SP_{int}}{gl_{int}}.$$

8. Cu rezultatele obținute se completează așa-numitul tabel ANOVA în care se va găsi și statistica testului (F):

Sursa de variație	SP	gl	$\overline{SP}$	F
Externă (între probe)	SP_{ext}	gl_{ext}	$\overline{SP}_{ext}$	$\overline{SP}_{ext}/\overline{SP}_{int}$
Internă (în probe)	SP_{int}	gl_{int}	$\overline{SP}_{int}$	
Totală	SP_t	gl_t		

Condiția testului constă în compararea statisticii F cu o valoare critică a distribuției Snedecor-Fisher tabelată în funcție de α, gradele de libertate externe (gl_{ext}) și gradele de libertate interne (gl_{int}) (anexa 2).

Dacă $F \geq F_{(\alpha,k-1,n_t-k)} \Rightarrow H_0$ se respinge și se acceptă H_1 pentru o probabilitate $p = 1 - \alpha$.

Valoarea critică, precum și probabilitatea statisticii testului (adică probabilitatea ca ipoteza nulă să fie adevărată) se pot calcula (anexa 3). Când se calculează probabilitatea asociată valorii F, atunci ipoteza nulă se respinge dacă $p(H_0)$ este mai mic de nivelul α și se va accepta ipoteza alternativă pentru o probabilitate de $1 - p(H_0)$.

În cazul în care la sfârșitul testului se respinge ipoteza nulă și se acceptă în consecință că există o diferență semnificativă, atunci analiza poate continua în sensul detectării diferențelor semnificative dintre mediile tuturor combinațiilor de probe. O evaluare rapidă a diferențelor dintre mediile perechilor de probe constă în inspectarea sau compararea grafică a intervalelor de confidență a mediilor populațiilor (secțiunea 5.1, fig. 5.5) din care au fost extrase probele. În general, există șanse mari ca diferența dintre două probe să fie semnificativă, dacă intervalele de confidență corespunzătoare nu se suprapun.

O modalitate mai sensibilă de detectare a diferențelor semnificative dintre medii este reprezentată de **testul Tukey**.

În cadrul acestui test se calculează diferențele în modul dintre mediile tuturor perechilor unice de probe:

Media Probei	$\bar{x}_2$	$\bar{x}_3$	...	$\bar{x}_k$
$\bar{x}_1$	$\lvert\bar{x}_1 - \bar{x}_2\rvert$	$\lvert\bar{x}_1 - \bar{x}_3\rvert$	...	$\lvert\bar{x}_1 - \bar{x}_k\rvert$
$\bar{x}_2$		$\lvert\bar{x}_2 - \bar{x}_3\rvert$	...	$\lvert\bar{x}_2 - \bar{x}_k\rvert$
...			...	...
$\bar{x}_{k-1}$				$\lvert\bar{x}_{k-1} - \bar{x}_k\rvert$

Se calculează apoi pentru fiecare pereche de probe câte o statistică T a testului, pornind de la o valoare critică Tukey ($q_{(\alpha,k,n_t-k)}$, anexa 2), suma de pătrate medie internă ($\overline{SP}_{int}$) și o medie armonică dintre dimensiunile probelor din fiecare pereche (n_h). Dacă probele au aceeași dimensiune, în loc de n_h se ia n.

$$n_h = \frac{2n_1n_2}{n_1+n_2}$$

$$T_{1,2} = q_{(\alpha,k,n_t-k)} \cdot \sqrt{\frac{\overline{SP}_{int}}{n_h}}$$

Dacă $|\bar{x}_1 - \bar{x}_2| \geq T_{1,2} \Rightarrow$ diferența dintre mediile populațiilor din care au fost extrase probele 1 și 2 este semnificativă pentru $p = 1 - \alpha$.

Dacă probele au aceeași dimensiune ($n_1 = n_2 = \cdots = n_k$), atunci diferențierea probelor se poate face prin investigarea sau compararea grafică a suprapunerii intervalelor de confidență a mediilor fiecărei probe. Limitele intervalului de confidență a unei medii rezultă din adunarea și scăderea la aceasta a statisticii Ţukey împărțită la doi:

$$\bar{x}_i \pm \frac{T}{2}.$$

Exemplul 8.2. Pe baza datelor din exemplul 8.1 să se afle dacă cele patru probe sunt diferite semnificativ.

Datele din exemplul 8.1 sunt distribuite în probe după un singur factor (zona), deci se realizează ANOVA unifactorială.

Ipotezele testului sunt:

H_0: mediile probelor nu diferă semnificativ;
H_1:mediile probelor diferă semnificativ.

Se calculează suma pătratelor tuturor valorilor prin adunarea sumelor pătratelor valorilor din fiecare probă:

$$\sum x_t^2 = 7952 + 9280 + 19700 + 39061 = 75993 .$$

Se calculează pătratul sumei totale prin adunarea sumelor valorilor din fiecare probă și ridicarea sumei la pătrat:

$$(\sum x_t)^2 = (292 + 308 + 464 + 667)^2 = (1731)^2 = 2996361 .$$

Se calculează numărul total de valori din toate probele:

$$n_t = 12 + 12 + 12 + 12 = 48 .$$

Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărei probe:

$$\sum \left[\frac{(\sum x_i)^2}{n_i}\right] = \frac{(292)^2}{12} + \frac{(308)^2}{12} + \frac{(464)^2}{12} + \frac{(667)^2}{12} = 70026{,}083 .$$

Se calculează cele trei sume de pătrate:

$$SP_t = 75993 - \frac{2996361}{48} = 13568{,}813$$
$$SP_{ext} = 70026{,}083 - \frac{2996361}{48} = 7601{,}896$$
$$SP_{int} = 13568{,}813 - 7601{,}896 = 5966{,}917 .$$

Se calculează cele trei grade de libertate:

$$gl_t = 48 - 1 = 47 \quad gl_{ext} = 4 - 1 = 3 \quad gl_{int} = 47 - 3 = 44 .$$

Se calculează sumele de pătrate medii:

$$\overline{SP}_{ext} = \frac{7601,896}{3} = 2533,965 \qquad \overline{SP}_{int} = \frac{5966,917}{44} = 135,612 .$$

Cu cele două sume de pătrate medii se calculează statistica testului:

$$F = \frac{2533,965}{135,612} = 18,685 .$$

Se completează tabelul ANOVA:

Sursa de variație	SP	gl	$\overline{SP}$	F
Externă	7601,896	3	2533,965	18,685
Internă	5966,917	44	135,612	
Totală	13568,813	47		

Se află valoarea critică tabelată (anexa 2) sau se calculează în funcție de nivelul de confidență, de gradele de libertate externe și de gradele de libertate interne (anexa 3):

$$F_{(0,05,3,44)} = 2,816 .$$

Statistica testului este mai mare decât valoarea critică și se poate respinge ipoteza nulă. Deci se acceptă că mediile probelor diferă semnificativ.

Probabilitatea ca ipoteza nulă să fie adevărată este $5,8 \cdot 10^{-8}$, adică extrem de mică.

În continuare, pentru a compara probele două câte două se folosește testul Tukey, pentru care trebuie calculate modulele diferențelor dintre perechile de probe:

Media Probei	**25,66**	**38,66**	**55,58**
24,33	1,33	14,33	31,25
25,66		13	29,92
38,66			16,92

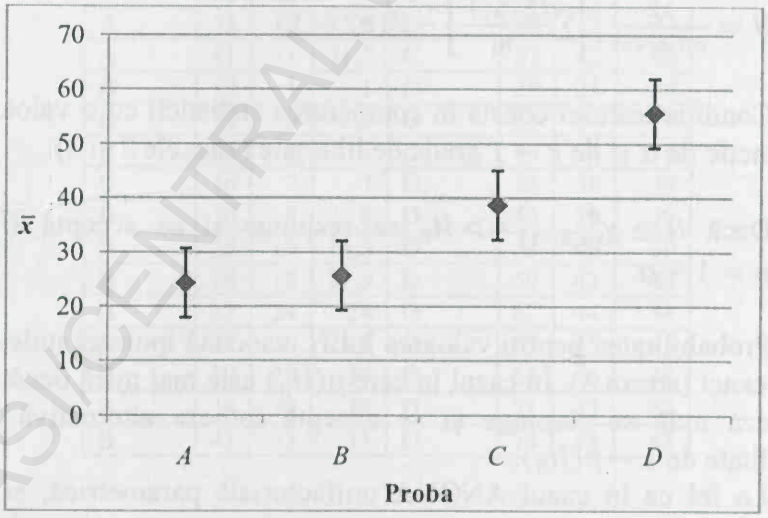
Se calculează statistica testului pornind de la valoarea critică Tukey (anexa 2):

$$q_{(0,05,4,44)} = 3,79$$

$$T = 3,79 \cdot \sqrt{\frac{135,612}{12}} = 12,74 .$$

Probele pentru care diferențele sunt mai mari decât statistica testului sunt diferite semnificativ: proba A diferă semnificativ de probele C și D, proba B diferă semnificativ de probele C și D și proba C diferă semnificativ de proba D. Probele A și B nu diferă semnificativ.

Pentru că probele au același număr de valori, se poate reprezenta grafic intervalul de confidență pentru fiecare probă adunând și scăzând din valoarea fiecărei medii statistica testului împărțită la doi ($\frac{T}{2} = \frac{12,74}{2} = 6,37$). Modul în care aceste intervale se suprapun arată care sunt probele diferite semnificativ.



În grafic se observă cum intervalele probelor A și B se suprapun, în timp ce intervalele celorlalte probe sunt separate pe verticală.

8.2.2. ANOVA unifactorială neparametrică Kruskal-Wallis

Acest tip de ANOVA este o alternativă neparametrică a modelului unifactorial. Ca urmare, se utilizează atunci când datele nu respectă condițiile de aplicare ale modelului unifactorial parametric.

Aranjarea datelor și ipotezele sunt similare cu cele de la ANOVA unifactorială parametrică.

Valorile din probe primesc ranguri împreună, la fel ca în cazul testului Mann-Whitney (secțiunea 7.1.2).

Se calculează apoi suma rapoartelor dintre pătratele sumelor rangurilor și numărul valorilor pentru fiecare probă:

$$\sum \frac{(\sum R_{x_i})^2}{n_i} = \frac{(\sum R_{x_1})^2}{n_1} + \frac{(\sum R_{x_2})^2}{n_2} + \cdots + \frac{(\sum R_{x_k})^2}{n_k}.$$

Statistica testului H se calculează pornind de la valoarea prezentată anterior și numărul total de valori din toate probele (n_t).

$$H = \frac{12}{n_t(n_t+1)} \cdot \left[\sum \frac{(\sum R_{x_i})^2}{n_i}\right] - 3(n_t + 1)$$

Condiția testului constă în compararea statisticii cu o valoare critică χ^2 în funcție de α și de $k-1$ grade de libertate (anexele 2 și 3).

Dacă $H \geq \chi^2_{(\alpha,k-1)} \Rightarrow H_0$ se respinge și se acceptă H_1 pentru $p = 1 - \alpha$.

Probabilitatea pentru valoarea lui H, asociată ipotezei nule, se poate calcula exact (anexa 3). În cazul în care $p(H_0)$ este mai mică decât valoarea α, ipoteza nulă se respinge și se acceptă ipoteza alternativă pentru o probabilitate de $1 - p(H_0)$.

La fel ca în cazul ANOVA unifactorială parametrică, și în cazul ANOVA unifactorială neparametrică se pot face comparații multiple dacă diferențele testate sunt semnificative. Pentru a compara probele două câte două se poate face pentru fiecare pereche câte un test U Mann-Whitney (secțiunea 7.1.2).

Exemplul 8.3. Pe baza datelor din exemplul 8.1 să se afle dacă cele patru probe sunt diferite semnificativ, considerând că nu sunt îndeplinite condițiile pentru a realiza ANOVA unifactorială parametrică.

În acest caz, se folosește alternativa neparametrică a ANOVA unifactorială, adică ANOVA Kruskal-Wallis.

Ipotezele sunt similare cu cele de la exemplul 8.2.

Se dau ranguri valorilor din cele patru probe împreună:

Proba	x	R_i	R_x	Proba	x	R_i	R_x
A	12	2	2,5	C	21	10	9,5
A	13	4	4,5	C	23	12	12
A	13	5	4,5	C	25	16	14,5
A	21	9	9,5	C	29	20	20
A	22	11	11	C	31	23	23
A	25	13	14,5	C	39	30	30
A	25	14	14,5	C	41	32	31,5
A	28	17	17,5	C	46	34	34,5
A	29	19	20	C	46	35	34,5
A	30	22	22	C	52	37	37
A	33	25	25,5	C	54	41	41
A	41	31	31,5	C	57	42	42
B	10	1	1	D	29	21	20
B	12	3	2,5	D	35	27	27
B	14	6	6	D	49	36	36
B	16	7	7	D	53	38	39
B	20	8	8	D	53	39	39
B	25	15	14,5	D	53	40	39
B	28	18	17,5	D	59	43	43
B	32	24	24	D	62	44	44
B	33	26	25,5	D	63	45	45
B	37	28	28	D	66	46	46
B	38	29	29	D	71	47	47
B	43	33	33	D	74	48	48

Se calculează apoi suma rapoartelor dintre pătratele sumelor rangurilor și numărul valorilor pentru fiecare probă:

$$\sum \frac{(\sum R_{x_l})^2}{n_l} = \frac{(177,5)^2}{12} + \frac{(196)^2}{12} + \frac{(329,5)^2}{12} + \frac{(473)^2}{12} = 33518,458 .$$

Se calculează statistica testului:

$$H = \frac{12}{48(48+1)} \cdot 33518,458 - 3(48+1) = 24,013 .$$

Se află valoarea critică pin căutare în tabel (anexa 2) sau prin calculare (anexa 3).

$$\chi^2_{(0,05,3)} = 7,815 .$$

Se verifică condiția testului:

$24,013 > 7,815 \Rightarrow H_0$ se respinge și se acceptă H_1 pentru $p = 1 - 0,05 = 0,95$.

Probabilitatea ca ipoteza nulă să fie adevărată (anexa 3) este de $2,5 \cdot 10^{-5}$, adică extrem de mică.

Concluzia testului este că probele diferă semnificativ unele de altele.

8.2.3. ANOVA bifactorială fără replicare

Acest model se mai numește și model cu o singură observație în celulă sau model cu observații repetate.

În cazul acestui model datele sunt grupate în probe în funcție de doi factori: C, care determină aranjarea datelor în c coloane ($C_i = C_1 \dots C_c$) și R, care determină aranjarea datelor în r rânduri sau linii ($R_j = R_1 \dots R_r$) (tab. 8.2). Între cei doi factori nu există interacțiune (secțiunea 8.2.5).

Tabelul 8.2. Distribuția valorilor în probe în ANOVA bifactorială

		C_i			
		C_1	C_2	…	C_c
R_j	R_1	$x_{c_1r_1}$	$x_{c_2r_1}$	…	$x_{c_cr_1}$
	R_2	$x_{c_1r_2}$	$x_{c_2r_2}$	…	$x_{c_cr_2}$
	…	…	…	…	…
	R_r	$x_{c_1r_r}$	$x_{c_2r_r}$	…	$x_{c_cr_r}$

Denumirea de model cu o singură observație în celulă vine de la faptul că, într-o celulă a tabelului, adică la intersecția unui nivel al unui factor cu un nivel al celuilalt factor (C_i, R_j) sau, mai simplu, la intersecția unui rând cu o coloană, se găsește o singură valoare (x_{c_i,r_j}).

Denumirea de model cu observații repetate provine de la un caz particular în care rândurile sunt reprezentate de unități de probă asupra cărora se fac observații repetate. O astfel de situație este similară cu cea în care se compară două probe neindependente cu ajutorul testului Student pentru perechi de observații (secțiunea 7.2.1), cu deosebirea că în cazul ANOVA se compară efectul a trei sau mai multe tratamente.

Conform acestui model orice observație poate fi definită ca:

$$\text{Obs.} = \begin{matrix}\text{Media}\\ \text{generală}\end{matrix} + \text{Efectul } C_i + \text{Efectul } R_j + \begin{matrix}\text{Eroarea}\\ \text{întâmplătoare.}\end{matrix}$$

Efectul nivelurilor factorilor dă variabilitatea externă (dintre probe), iar eroarea întâmplătoare este rezultatul variabilității interne (din cadrul probelor).

În cazul acestui model, variabilitatea externă (s^2_{ext}) poate fi la rândul ei descompusă în variabilitatea dintre coloane și variabilitatea dintre rânduri $(s_c^2 + s_r^2)$, astfel că relația de principiu a ANOVA bifactorială devine:

$$s_t^2 = s_c^2 + s_r^2 + s_{int}^2 .$$

Dacă varianțele sau sumele de pătrate medii $(\overline{SP})$ le considerăm rapoarte între sumele de pătrate (SP) și grade de libertate (gl), atunci:

$$SP_t = SP_c + SP_r + SP_{int}$$

$$gl_t = gl_c + gl_r + gl_{int} .$$

Ipotezele care se testează prin intermediul acestui model pot viza semnificația efectului unui singur factor sau a ambilor factori, caz în care se vor emite două ipoteze nule și două ipoteze alternative pentru același test. Pentru date experimentare ipotezele sunt formulate astfel:

H_{0_1}: nu există diferențe semnificative între efectele nivelurilor primului factor (C_i) asupra variabilei;

H_{0_2}: nu există diferențe semnificative între efectele nivelurilor celui de-al doilea factor (R_j) asupra variabilei;

H_{1_1}: diferențele dintre efectele nivelurilor primului factor (C_i) asupra variabilei sunt semnificative;

H_{1_2}: diferențele dintre efectele nivelurilor celui de-al doilea factor (R_j) asupra variabilei sunt semnificative.

În cazul unor date obținute în urma realizării unor observații efectuate asupra unei variabile în populații diferite, ipotezele pot fi formulate astfel:

H_{0_1}: mediile populațiilor corespunzătoare nivelurilor primului factor (C_i) nu diferă semnificativ;

H_{0_2}: mediile populațiilor corespunzătoare nivelurilor celui de-al doilea factor (R_j) nu diferă semnificativ;

H_{1_1}: mediile populațiilor corespunzătoare nivelurilor primului factor (C_i) diferă semnificativ;

H_{1_2}: mediile populațiilor corespunzătoare nivelurilor primului factor (R_j) diferă semnificativ.

Dacă se aplică ANOVA unifactorială pentru c coloane și r rânduri, trebuie parcurse următoarele etape:

1. Se calculează suma pătratelor tuturor valorilor ($\sum x_t^2$) prin adunarea sumelor pătratelor valorilor pe coloane sau pe rânduri ($\sum x_{c_i}^2$ sau $\sum x_{r_j}^2$):

$$\sum x_t^2 = \sum x_{c_1}^2 + \sum x_{c_2}^2 + \cdots + \sum x_{c_c}^2$$

Sau

$$\sum x_t^2 = \sum x_{r_1}^2 + \sum x_{r_2}^2 + \cdots + \sum x_{r_r}^2 .$$

2. Se calculează pătratul sumei totale ($(\sum x_t)^2$) prin adunarea sumelor valorilor din fiecare coloană ($\sum x_{c_i}$) sau rând ($\sum x_{r_j}$), urmată de ridicarea la pătrat:

$$(\sum x_t)^2 = \left(\sum x_{c_1} + \sum x_{c_2} + \cdots + \sum x_{c_c}\right)^2$$

sau

$$(\sum x_t)^2 = \left(\sum x_{r_1} + \sum x_{r_2} + \cdots + \sum x_{r_r}\right)^2 .$$

3. Se calculează numărul total de valori din toate probele (n_t) prin însumarea dimensiunilor coloanelor (n_{c_i}) sau rândurilor (n_{r_j}):

$$n_t = n_{c_1} + n_{c_2} + \cdots + n_{c_c}$$

sau

$$n_t = n_{r_1} + n_{r_2} + \cdots + n_{r_r} .$$

4. Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărei coloane:

$$\sum \left[\frac{\left(\sum x_{c_i}\right)^2}{n_{c_i}}\right] = \frac{\left(\sum x_{c_1}\right)^2}{n_{c_1}} + \frac{\left(\sum x_{c_2}\right)^2}{n_{c_2}} + \cdots + \frac{\left(\sum x_{c_c}\right)^2}{n_{c_c}} .$$

5. Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărui rând:

$$\sum \left[\frac{\left(\sum x_{r_j}\right)^2}{n_{r_j}}\right] = \frac{\left(\sum x_{r_1}\right)^2}{n_{r_1}} + \frac{\left(\sum x_{r_2}\right)^2}{n_{r_2}} + \cdots + \frac{\left(\sum x_{r_r}\right)^2}{n_{r_r}} .$$

6. Se calculează sumele de pătrate: totală (SP_t), dintre coloane (SP_c), dintre rânduri (SP_r) și cea internă (SP_{int}):

$$SP_t = \sum x_t^2 - \frac{(\sum x_t)^2}{n_t} \qquad SP_c = \sum \left[\frac{\left(\sum x_{c_i}\right)^2}{n_{c_i}}\right] - \frac{(\sum x_t)^2}{n_t}$$

$$SP_r = \sum \left[\frac{\left(\sum x_{r_j}\right)^2}{n_{r_j}}\right] - \frac{(\sum x_t)^2}{n_t} \qquad SP_{int} = SP_t - (SP_c + SP_r) .$$

7. Se calculează numărul gradelor de libertate totale (gl_t), pentru coloane (gl_c), pentru rânduri (gl_r) și cele interne (gl_{int}):

$$gl_t = n_t - 1$$
$$gl_c = c - 1$$
$$gl_r = r - 1$$
$$gl_{int} = gl_t - (gl_c + gl_r) = n_t - c - r + 1 = (c-1)(r-1)\,.$$

8. Sumele de pătrate medii ($\overline{SP}$) se calculează împărțind sumele de pătrate (SP) la gradele de libertate corespunzătoare (gl):

$$\overline{SP}_c = \frac{SP_c}{gl_c}$$
$$\overline{SP}_r = \frac{SP_r}{gl_r}$$
$$\overline{SP}_{int} = \frac{SP_{int}}{n_{int}}\,.$$

9. Cu rezultatele obținute se completează așa-numitul tabel ANOVA în care se va găsi și statistica testului (F):

Sursa de variație	SP	gl	$\overline{SP}$	F
Externă, între coloane	SP_c	gl_c	$\overline{SP}_c$	$F_c = \overline{SP}_c/\overline{SP}_{int}$
Externă, între rânduri	SP_r	gl_r	$\overline{SP}_r$	$F_r = \overline{SP}_r/\overline{SP}_{int}$
Internă	SP_{int}	gl_{int}	$\overline{SP}_{int}$	
Totală	SP_t	gl_t		

Condiția testului constă în compararea statisticii F cu o valoare critică tabelată în funcție de α, gradele de libertate externe (gl_{ext}) și gradele de libertate interne (gl_{int}) (anexa 2).

Dacă $F_c \geq F_{(\alpha,c-1,gl_{int})} \Rightarrow H_{0_1}$ se respinge și se acceptă H_{1_1} pentru o probabilitate $p = 1 - \alpha$.
Dacă $F_r \geq F_{(\alpha,r-1,gl_{int})} \Rightarrow H_{0_2}$ se respinge și se acceptă H_{1_2} pentru o probabilitate $p = 1 - \alpha$.

Valoarea critică, precum și probabilitatea statisticii testului (adică

probabilitatea ca ipoteza nulă să fie adevărată) se pot calcula (anexa 3). Când se calculează probabilitatea asociată valorii F, atunci ipoteza nulă se respinge dacă $p(H_0)$ este mai mic de nivelul α și se va accepta ipoteza alternativă pentru o probabilitate de $1 - p(H_0)$.

Exemplul 8.4. S-a urmărit densitatea realizată de o anumită specie de plantă în cinci suprafețe de probă (de la *A* la *E*) urmărite timp de patru ani (de la *I* la *IV*) dintr-o zonă supusă reconstrucției ecologice. A crescut semnificativ densitatea plantei în timp?

		An			
		I	*II*	*III*	*IV*
Proba	*A*	3	7	10	18
	B	8	10	15	20
	C	17	30	33	65
	D	8	11	31	60
	E	2	3	17	15

	I	*II*	*III*	*IV*
A	1,0986	1,9459	2,3026	2,8904
B	2,0794	2,3026	2,7081	2,9957
C	2,8332	3,4012	3,4965	4,1744
D	2,0794	2,3979	3,4340	4,0943
E	0,6931	1,0986	2,8332	2,7081

Se poate observa că datele sunt organizate după doi factori: an și probă. Deci se poate realiza un model bifactorial de ANOVA. Dat fiind faptul că în fiecare celulă există o singură valoare, se va realiza ANOVA bifactorială fără replicare.

Fiind vorba de relativ puține date ce reprezintă număr de entități, se poate face o transformare care să normalizeze distribuția și să stabilizeze varianța probelor.

Întrebarea problemei se referă doar la diferențele semnificative dintre ani (coloane). Dacă s-ar fi dorit și evidențierea diferențelor dintre suprafețele de probă (rânduri), atunci trebuie testate două seturi de ipoteze: una nulă și una alternativă pentru fiecare factor:

H_{0_1}: nu există diferențe semnificative între densități pe ani;
H_{0_2}: nu există diferențe semnificative între densități pe probe;
H_{1_1}: diferențele dintre densitățile pe ani sunt semnificative;
H_{1_2}: diferențele dintre densitățile pe probe sunt semnificative.

Se calculează suma pătratelor tuturor valorilor logaritmate:

$$\sum x_t^2 = 18{,}363 + 27{,}613 + 44{,}680 + 58{,}851 = 149{,}508 .$$

Se calculează pătratul sumei totale la pătrat:

$$(\sum x_t)^2 = (8{,}784 + 11{,}146 + 14{,}774 + 16{,}863)^2 = (51{,}567)^2 = \\ = 2659{,}185 .$$

Se calculează numărul total de valori din toate probele

$$n_t = 5 + 5 + 5 + 5 = 20 .$$

Se calculează suma rapoartelor dintre pătratul sumei şi dimensiunea fiecărei coloane:

$$\sum \left[\frac{\left(\sum x_{c_i}\right)^2}{n_{c_i}} \right] = \frac{(8{,}784)^2}{5} + \frac{(11{,}146)^2}{5} + \frac{(14{,}774)^2}{5} + \frac{(16{,}863)^2}{5} = 140{,}806 .$$

Se calculează suma rapoartelor dintre pătratul sumei şi dimensiunea fiecărui rând:

$$\sum \left[\frac{\left(\sum x_{r_j}\right)^2}{n_{r_j}} \right] = \frac{(8{,}237)^2}{4} + \frac{(10{,}086)^2}{4} + \frac{(13{,}905)^2}{4} + \frac{(12{,}006)^2}{4} + \frac{(7{,}333)^2}{4} = \\ = 140{,}212 .$$

Se calculează sumele de pătrate:

$$SP_t = 149{,}508 - \frac{(51{,}567)^2}{20} = 16{,}549$$
$$SP_c = 140{,}806 - \frac{(51{,}567)^2}{20} = 7{,}847$$
$$SP_r = 140{,}212 - \frac{(51{,}567)^2}{20} = 7{,}252$$
$$SP_{int} = 16{,}549 - (7{,}847 + 7{,}252) = 1{,}450 .$$

Se calculează numărul gradelor de libertate:

$gl_t = 20 - 1 = 19$ $gl_c = 4 - 1 = 3$
$gl_r = 5 - 1 = 4$ $gl_{int} = (4 - 1)(5 - 1) = 12$.

Se calculează sumele de pătrate medii:

$\overline{SP_c} = \frac{7,847}{3} = 2,616$ $\overline{SP_r} = \frac{7,252}{4} = 1,813$ $\overline{SP_{int}} = \frac{1,450}{12} = 0,121$.

Se calculează statisticile testului și se completează tabelul ANOVA:

$F_c = \frac{2,616}{0,121} = 21,620$ $F_r = \frac{1,813}{0,121} = 14,983$.

Sursa de variație	SP	gl	$\overline{SP}$	F
Externă, între coloane	7,847	3	2,616	21,620
Externă, între rânduri	7,252	4	1,813	14,983
Internă	1,450	12	0,121	
Totală	16,549	19		

Se află valorile critice pentru fiecare sursă de variație externă (anexa 2 sau 3):

$F_{(0,05,3,12)} = 3,490$ $F_{(0,05,4,12)} = 3,259$.

Se verifică condiția testului:

$21,620 > 3,490 \Rightarrow H_{0_1}$ se respinge și se acceptă H_{1_1} pentru o probabilitate $p = 1 - 0,05 = 0,95$.
$14,983 > 3,259 \Rightarrow H_{0_2}$ se respinge și se acceptă H_{1_2} pentru o probabilitate $p = 1 - 0,05 = 0,95$.

Probabilitățile celor două ipoteze nule se pot calcula (anexa 3):

$p(H_{0_1}) = 3,92 \cdot 10^{-5}$ și $p(H_{0_2}) = 1,18 \cdot 10^{-4}$.

Concluzia testului este că modificările induse de trecerea anilor densității plantei analizate sunt semnificative. De asemenea, există și o diferență semnificativă între suprafețele de probă din punctul de vedere al densității speciei analizate.

8.2.4. ANOVA bifactorială neparametrică Friedman

Se utilizează ca alternativă neparametrică a modelului bifactorial atunci când se fac observații repetate asupra acelorași unități de probă, obținându-se probe neindependente. Din acest punct de vedere ANOVA bifactorială neparametrică este similară testului Wilcoxon (secțiunea 7.2.2), doar că se folosește pentru trei sau mai multe probe.

Aranjarea datelor și ipotezele sunt similare cu cele de la ANOVA bifactorială parametrică, cu deosebirea că se testează doar ipotezele H_{0_1} și H_{1_1} referitoare la factorul ce determină aranjarea datelor în coloane.

Valorile fiecărui rând primesc ranguri în mod independent. Rangurile pentru fiecare rând se atribuie conform algoritmului prezentat în secțiunea 2.1, tab. 2.3.

Logica acestui test este că, dacă nu există diferențe între efectele nivelului primului factor (C_i) sau, mai simplu, dacă nu există diferențe între coloane, sumele rangurilor coloanelor ar trebui să fie aproximativ egale, deoarece orice rang are șanse egale să apară în orice coloană.

Se calculează suma rangurilor valorilor din fiecare coloană ($\sum R_{x_{c_i}}$), după care se ridică la pătrat:

$$\left(\sum R_{x_{c_i}}\right)^2 = \left(R_{x_{c_i r_1}} + R_{x_{c_i r_2}} + \cdots + R_{x_{c_i r_r}}\right)^2 .$$

Se calculează suma pătratelor sumelor rangurilor pe coloană:

$$\sum\left(\sum R_{x_{c_i}}\right)^2 = \left(\sum R_{x_{c_1}}\right)^2 + \left(\sum R_{x_{c_2}}\right)^2 + \cdots + \left(\sum R_{x_{c_c}}\right)^2 .$$

Statistica testului se calculează cu ajutorul formulei:

$$Q = \frac{12}{rc(c+1)} \cdot \left[\sum\left(\sum R_{x_{c_i}}\right)^2\right] - 3r(c+1) .$$

Condiția testului compară valoarea statisticii cu o valoare critică χ^2 tabelată (anexa 2) sau calculată (anexa 3) în funcție de α și de numărul de coloane minus unu grade de libertate $(c-1)$.

Dacă $Q \geq \chi^2_{(\alpha,c-1)} => H_{0_1}$ se respinge, H_{1_1} se acceptă pentru $p = 1 - \alpha$.

Probabilitatea pentru valoarea lui χ^2_r, asociată ipotezei nule, se poate calcula exact (anexa 3). În cazul în care $p(H_0)$ este mai mică de valoarea α, ipoteza nulă se respinge și se acceptă ipoteza alternativă pentru o probabilitate de $1 - p(H_0)$.

Deci acesta verifică doar setul de ipoteze care se referă la factorul după care se aranjează datele pe coloane. Cu alte cuvinte, dacă se respinge ipoteza nulă, concluzia este că există o diferență semnificativă între coloane.

Exemplul 8.5. Să se răspundă la întrebarea de la exemplul 8.4 considerând că nu sunt îndeplinite condițiile de aplicare a ANOVA bifactoriale fără replicare, parametrice.

În acest caz se apelează la alternativa neparametrică a acestui tip de ANOVA, adică la ANOVA Freidman.

Ipotezele testului sunt similare celor de la exemplul 8.4, cu deosebirea că ANOVA Friedman testează doar un singur set de ipoteze, cele care vizează coloanele. Pentru a testa și ipotezele referitoare la rânduri, datele trebuie rearanjate astfel încât coloanele să devină rânduri și rândurile coloane. Altfel spus, tabelul de date trebuie rotit.

Pentru a efectua testul trebuie ca valorile din fiecare rând să primească ranguri, după care se calculează suma rangurilor pe coloane:

	I	*II*	*III*	*IV*
A	1	2	3	4
B	1	2	3	4
C	1	2	3	4
D	1	2	3	4
E	1	2	4	3
$\sum R_{x_{c_i}}$	5	10	16	19

Se calculează apoi suma sumelor ridicate la pătrat ale rangurilor pe coloane:

$$\left[\sum\left(\sum R_{x_{ci}}\right)^2\right] = (5)^2 + (10)^2 + (16)^2 + (19)^2 = 742 .$$

Se calculează statistica testului:

$$Q = \frac{12}{5 \cdot 4(4+1)} \cdot 742 - 3 \cdot 5(4+1) = 14{,}04 .$$

Se află valoarea critică χ^2 (anexa 2 sau anexa 3): $\chi^2_{(0,05,3)} = 7{,}815$.

Statistica testului este mai mare decât valoarea critică, deci se respinge ipoteza nulă (H_{0_1}) și se acceptă ipoteza alternativă (H_{1_1}).

Probabilitatea ca ipoteza nulă să fie adevărată se poate calcula (anexa 3) și este egală cu 0,0029.

Concluzia testului este că probele (coloanele) diferă semnificativ.

8.2.5. ANOVA bifactorială cu replicare

Acest model se mai numește si model cu număr egal de observații în celulă.

La fel ca și la ANOVA bifactorială fără replicare, datele sunt grupate în probe în funcție de doi factori: C, care determină aranjarea datelor în c coloane ($C_i = C_1 \dots C_c$), și R, care determină aranjarea datelor în r rânduri sau linii ($R_j = R_1 \dots R_r$) (tab. 8.3). În cazul acestui model particularitatea de formă constă în faptul că într-o celulă a tabelului, adică la intersecția unui nivel al unui factor cu un nivel al celuilalt factor (C_i, R_j) sau, mai simplu, la intersecția unui rând cu o coloană, se găsesc mai multe valori ($x_{c_i,r_j,n}$), adică n observații replicate.

Particularitatea de principiu a acestui model este reprezentată de faptul că presupune o interacțiune între cei doi factori. Practic, interacțiunea este evidentă dacă efectul unor anumite niveluri ale factorilor asupra variabilei investigate se modifică într-o manieră neaditivă.

Tabelul 8.3. Distribuția valorilor în probe în ANOVA bifactorială

		C_i			
		C_1	C_2	...	C_c
R_j	R_1	$x_{c_1r_11}$ ... $x_{c_1r_1n}$	$x_{c_2r_11}$ ... $x_{c_2r_1n}$	...	$x_{c_cr_11}$ ... $x_{c_cr_1n}$
	R_2	$x_{c_1r_21}$ ... $x_{c_1r_2n}$	$x_{c_2r_21}$ ... $x_{c_2r_2n}$	...	$x_{c_cr_21}$ ... $x_{c_cr_2n}$
	...	...	...	...	...
	R_r	$x_{c_1r_r1}$ ... $x_{c_1r_rn}$	$x_{c_2r_r1}$ ... $x_{c_2r_rn}$	...	$x_{c_cr_r1}$ ... $x_{c_cr_rn}$

Evidențierea interacțiunii se poate realiza prin vizualizarea grafică a mediilor celulelor (a mediilor valorilor de la 1 la n din fiecare celulă). Dacă liniile care unesc mediile după un factor sunt mai mult sau mai puțin paralele, atunci între factori nu există interacțiune. În figura 8.1 se poate observa că trecerea de la un nivel al factorului C la altul (trecerea de la C_1 la C_2) modifică variabila în sensul descreșterii mediilor cu aceeași valoare și pentru rândul 1, și pentru rândul 2. Modificarea variabilei se realizează în același sens și sub acțiunea nivelurilor factorului R. Deci, în acest caz, factorii C și R au o acțiune aditivă negativă (de scădere) a variabilei investigate.

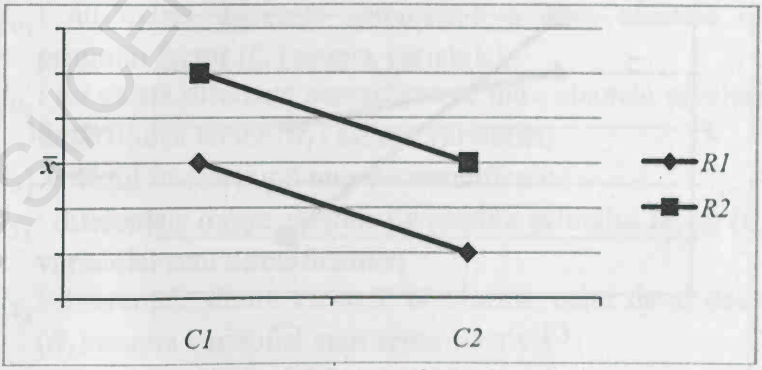


Figura 8.1. Interacțiune absentă

Dacă liniile ce unesc mediile după un factor sunt evident neparalele, înseamnă că între cei doi factori există o interacțiune sau efectele factorilor asupra variabilei sunt neaditive. În figura 8.2, efectele factorului C asupra variabilei modifică efectele factorului R în sensul că pentru R_1 efectul trecerii de la C_1 la C_2 constă într-o scădere mai mare decât pentru R_2.

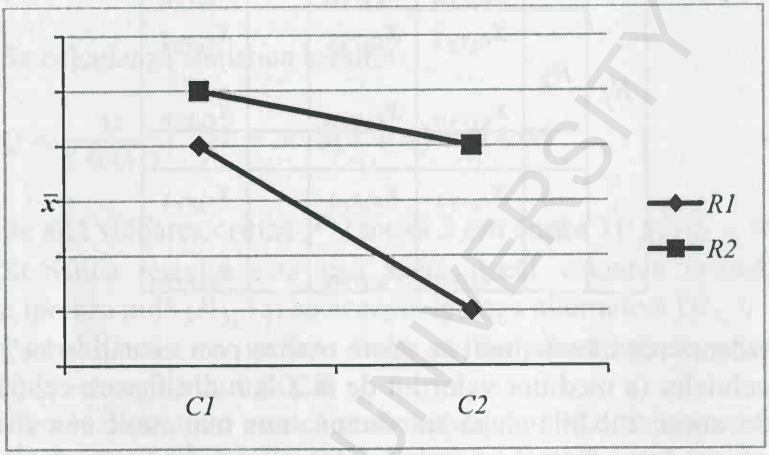


Figura 8.2. Interacțiune: C_2R_1

În figura 8.3 este vorba tot de interacțiune între factori, dar în acest caz interacțiunea constă într-o descreștere mai amplă în cazul R_2 sub acțiunea factorului C.

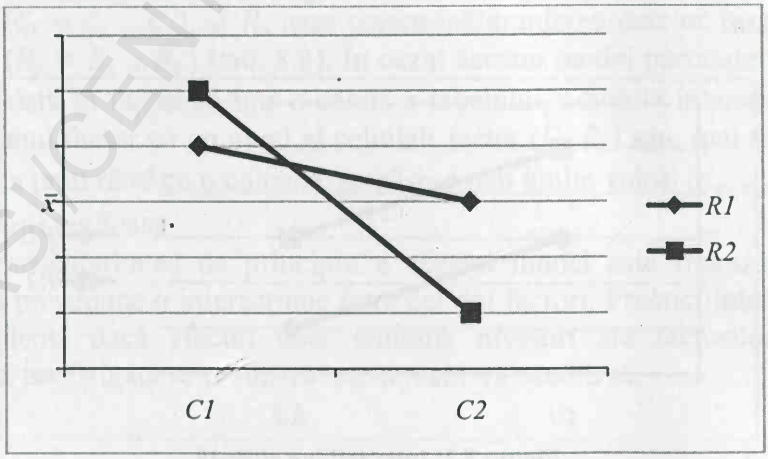


Figura 8.3. Interacțiune: C_2R_2

Conform acestui model orice observație poate fi definită ca:

$$\text{Obs.} = \underset{\text{generală}}{\text{Media}} + \underset{C_i}{\text{Efect}} + \underset{R_j}{\text{Efect}} + \underset{C_iR_j}{\text{Efect Interacțiune}} + \underset{\text{întâmplătoare.}}{\text{Eroarea}}$$

Efectul nivelurilor factorilor dă variabilitatea externă (dintre probe), iar eroarea întâmplătoare este rezultatul variabilității interne (din cadrul probelor).

În cazul acestui model variabilitatea externă (s^2_{ext}) poate fi rândul ei descompusă în variabilitatea dintre coloane, variabilitatea dintre rânduri și variabilitatea cauzată de interacțiune ($s^2_c + s^2_r + s^2_i$), astfel că relația de principiu a ANOVA bifactorială devine:

$$s^2_t = s^2_c + s^2_r + s^2_i + s^2_{int} .$$

Dacă varianțele sau sumele de pătrate medii ($\overline{SP}$) le considerăm rapoarte între sumele de pătrate (SP) și grade de libertate (gl), atunci:

$$SP_t = SP_c + SP_r + SP_i + SP_{int}$$
$$gl_t = gl_c + gl_r + gl_i + gl_{int} .$$

Ipotezele care se testează prin intermediul acestui model pot viza semnificația efectului unui singur factor sau a ambilor factori, caz în care se vor emite două ipoteze nule și două ipoteze alternative pentru același test. Pentru date experimentale ipotezele sunt formulate astfel:

H_{0_1}: nu există diferențe semnificative între efectele nivelurilor primului factor (C_i) asupra variabilei;

H_{0_2}: nu există diferențe semnificative între efectele nivelurilor celui de-al doilea factor (R_j) asupra variabilei;

H_{0_3}: efectul interacțiunii nu este semnificativ;

H_{1_1}: diferențele dintre efectele nivelurilor primului factor (C_i) asupra variabilei sunt semnificative;

H_{1_2}: diferențele dintre efectele nivelurilor celui de-al doilea factor (R_j) asupra variabilei sunt semnificative;

H_{1_3}: efectul interacțiunii este semnificativ.

În cazul unor date obținute în urma realizării unor observații efectuate asupra unei variabile în populații diferite, ipotezele pot fi formulate astfel:

H_{0_1}: mediile populațiilor corespunzătoare nivelurilor primului factor (C_i) nu diferă semnificativ;

H_{0_2}: mediile populațiilor corespunzătoare nivelurilor celui de-al doilea factor (R_j) nu diferă semnificativ;

H_{0_3}: mediile corespunzătoare interacțiunii nivelurilor factorilor (C_i, R_j) nu diferă semnificativ;

H_{1_1}: mediile populațiilor corespunzătoare nivelurilor primului factor (C_i) diferă semnificativ;

H_{1_2}: mediile populațiilor corespunzătoare nivelurilor primului factor (R_j) diferă semnificativ;

H_{1_3}: mediile corespunzătoare interacțiunii nivelurilor factorilor (C_i, R_j) diferă semnificativ.

Dacă se aplică ANOVA unifactorială pentru c coloane, r rânduri și k valori în fiecare celulă, trebuie parcurse următoarele etape:

1. Se calculează suma pătratelor tuturor valorilor $(\sum x_t^2)$ prin adunarea pătratelor tuturor valorilor pe coloane sau pe rânduri $(\sum x_{c_i}^2$ sau $\sum x_{r_j}^2)$:

$$\sum x_t^2 = \sum x_{c_1}^2 + \sum x_{c_2}^2 + \cdots + \sum x_{c_c}^2$$

sau

$$\sum x_t^2 = \sum x_{r_1}^2 + \sum x_{r_2}^2 + \cdots + \sum x_{r_r}^2 .$$

2. Se calculează pătratul sumei totale $((\sum x_t)^2)$ prin adunarea sumelor valorilor din fiecare coloană $(\sum x_{c_i})$ sau rând $(\sum x_{r_j})$, urmată de ridicarea la pătrat:

$$(\sum x_t)^2 = \left(\sum x_{c_1} + \sum x_{c_2} + \cdots + \sum x_{c_c}\right)^2$$

sau

$$(\sum x_t)^2 = \left(\sum x_{r_1} + \sum x_{r_2} + \cdots + \sum x_{r_r}\right)^2 .$$

3. Se calculează numărul total de valori din toate probele (n_t) prin însumarea dimensiunilor coloanelor (n_{c_i}) sau rândurilor (n_{r_j}) sau prin înmulțirea numărului de valori din celulă cu numărul celulelor egal cu produsul dintre numărul coloanelor și rândurilor:

$$n_t = n_{c_1} + n_{c_2} + \cdots + n_{c_c}$$

sau

$$n_t = n_{r_1} + n_{r_2} + \cdots + n_{r_r}$$

sau

$$n_t = n \cdot c \cdot r\ .$$

4. Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărei coloane:

$$\sum\left[\frac{\left(\sum x_{c_i}\right)^2}{n_{c_i}}\right] = \frac{\left(\sum x_{c_1}\right)^2}{n_{c_1}} + \frac{\left(\sum x_{c_2}\right)^2}{n_{c_2}} + \cdots + \frac{\left(\sum x_{c_c}\right)^2}{n_{c_c}}\ .$$

5. Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărui rând:

$$\sum\left[\frac{\left(\sum x_{r_j}\right)^2}{n_{r_j}}\right] = \frac{\left(\sum x_{r_1}\right)^2}{n_{r_1}} + \frac{\left(\sum x_{r_2}\right)^2}{n_{r_2}} + \cdots + \frac{\left(\sum x_{r_r}\right)^2}{n_{r_r}}\ .$$

6. Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărei celule ($n_{c_i r_j} = n$):

$$\sum\left[\frac{\left(\sum x_{c_i r_j}\right)^2}{n_{c_i r_j}}\right] = \frac{\left(\sum x_{c_1 r_1}\right)^2}{n_{c_1 r_1}} + \frac{\left(\sum x_{c_1 r_2}\right)^2}{n_{c_1 r_2}} + \cdots + \frac{\left(\sum x_{c_c r_r}\right)^2}{n_{c_c r_r}}$$

sau

$$\sum\left[\frac{\left(\sum x_{c_i r_j}\right)^2}{n}\right] = \frac{\left(\sum x_{c_1 r_1}\right)^2}{n} + \frac{\left(\sum x_{c_1 r_2}\right)^2}{n} + \cdots + \frac{\left(\sum x_{c_c r_r}\right)^2}{n}\ .$$

7. Se calculează sumele de pătrate: totală (SP_t), dintre coloane (SP_c), dintre rânduri (SP_r), de interacțiune (SP_i) și cea internă (SP_{int}):

$$SP_t = \sum x_t^2 - \frac{(\sum x_t)^2}{n_t}$$

$$SP_c = \sum \left[\frac{\left(\sum x_{c_i}\right)^2}{n_{c_i}} \right] - \frac{(\sum x_t)^2}{n_t}$$

$$SP_r = \sum \left[\frac{\left(\sum x_{r_j}\right)^2}{n_{r_j}} \right] - \frac{(\sum x_t)^2}{n_t}$$

$$SP_i = \sum \left[\frac{\left(\sum x_{c_i r_j}\right)^2}{n_{c_i r_j}} \right] - \frac{(\sum x_t)^2}{n_t} - (SP_c + SP_r)$$

$$SP_{int} = SP_t - (SP_c + SP_r + SP_i) .$$

8. Se calculează numărul gradelor de libertate totale (gl_t) pentru coloane (gl_c), pentru rânduri (gl_r), pentru interacțiune (gl_i) și pentru cele interne (gl_{int}):

$$gl_t = n_t - 1$$
$$gl_c = c - 1$$
$$gl_r = r - 1$$
$$gl_i = gl_c \cdot gl_r = (c-1)(r-1)$$
$$gl_{int} = gl_t - (gl_c + gl_r + gl_i) = n_t - cr .$$

9. Sumele de pătrate medii ($\overline{SP}$) se calculează împărțind sumele de pătrate (SP) la gradele de libertate corespunzătoare (gl):

$$\overline{SP_c} = \frac{SP_c}{gl_c}$$
$$\overline{SP_r} = \frac{SP_r}{gl_r} \quad .$$
$$\overline{SP_i} = \frac{SP_i}{gl_i}$$
$$\overline{SP_{int}} = \frac{SP_{int}}{n_{int}} .$$

10. Cu rezultatele obținute se completează așa-numitul tabel ANOVA în care se va găsi și statistica testului (F):

Sursa de variație	SP	gl	$\overline{SP}$	F
Externă, între coloane	SP_c	gl_c	$\overline{SP_c}$	$F_c = \overline{SP_c}/\overline{SP_{int}}$
Externă, între rânduri	SP_r	gl_r	$\overline{SP_r}$	$F_r = \overline{SP_r}/\overline{SP_{int}}$
Externă, de interacțiune (între celule)	SP_i	gl_i	$\overline{SP_i}$	$F_i = \overline{SP_i}/\overline{SP_{int}}$
Internă	SP_{int}	gl_{int}	$\overline{SP_{int}}$	
Totală	SP_t	gl_t		

Condiția testului constă în compararea statisticii F cu o valoarea critică tabelată în funcție de α, gradele de libertate externe (gl_{ext}) și gradele de libertate interne (gl_{int}) (anexa 2 sau anexa 3).

Dacă $F_c \geq F_{(\alpha,c-1,gl_{int})} \Rightarrow H_{0_1}$ se respinge și se acceptă H_{1_1} pentru o probabilitate $p = 1 - \alpha$.
Dacă $F_r \geq F_{(\alpha,r-1,gl_{int})} \Rightarrow H_{0_2}$ se respinge și se acceptă H_{1_2} pentru o probabilitate $p = 1 - \alpha$.
Dacă $F_i \geq F_{(\alpha,gl_i,gl_{int})} \Rightarrow H_{0_3}$ se respinge și se acceptă H_{1_3} pentru o probabilitate $p = 1 - \alpha$.

Valoarea critică, precum și probabilitatea statisticii testului (adică probabilitatea ca ipoteza nulă să fie adevărată) se pot calcula (anexa 3). Când se calculează probabilitatea asociată valorii F, atunci ipoteza nulă se respinge dacă $p(H_0)$ este mai mic de nivelul α și se va accepta ipoteza alternativă pentru o probabilitate de $1 - p(H_0)$.

Dacă diferențele testate sunt semnificative, atunci se pot face comparații multiple. Pentru aceasta se utilizează testul Tukey, în cadrul căruia se calculează diferențele în modul dintre mediile tuturor perechilor unice de probe (celule):

Media probei	$\bar{x}_{c_2r_1}$	$\bar{x}_{c_3r_1}$	...	$\bar{x}_{c_cr_r}$
$\bar{x}_{c_1r_1}$	$\lvert\bar{x}_{c_1r_1} - \bar{x}_{c_2r_1}\rvert$	$\lvert\bar{x}_{c_1r_1} - \bar{x}_{c_3r_1}\rvert$	...	$\lvert\bar{x}_{c_1r_1} - \bar{x}_{c_cr_r}\rvert$
$\bar{x}_{c_2r_1}$		$\lvert\bar{x}_{c_2r_1} - \bar{x}_{c_3r_1}\rvert$	...	$\lvert\bar{x}_{c_2r_1} - \bar{x}_{c_cr_r}\rvert$
...			...	...
$\bar{x}_{c_{c-1}r_{r-1}}$				$\lvert\bar{x}_{c_{c-1}r_{r-1}} - \bar{x}_{c_cr_r}\rvert$

Se calculează apoi pentru statistica T a testului pornind de la o valoare critică Tukey ($q_{(\alpha,c\cdot r,n_t-c\cdot r)}$, anexa 2), suma de pătrate medie internă ($\overline{SP}_{int}$) și numărul valorilor dintr-o celulă (n).

$$T = q_{(\alpha,cr,n_t-cr)} \cdot \sqrt{\frac{\overline{SP}_{int}}{n}}$$

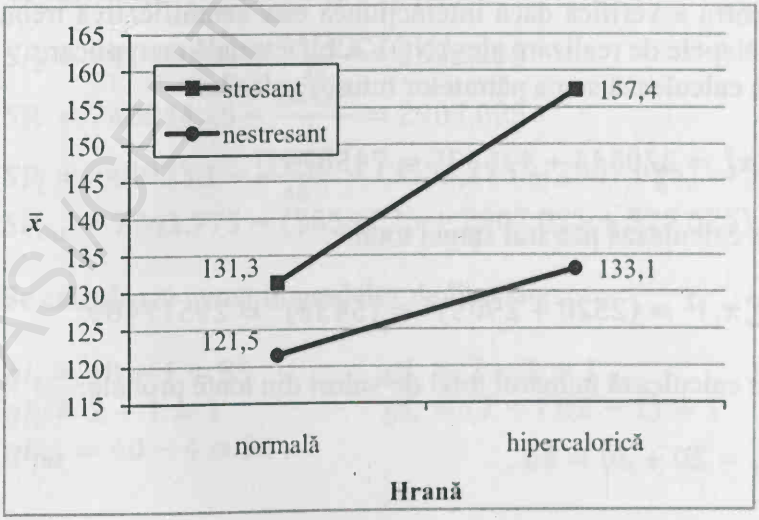
Dacă diferența absolută dintre oricare două medii $|\bar{x}_1 - \bar{x}_2| \geq T \Rightarrow$ diferența este semnificativă pentru $p = 1 - \alpha$.

Exemplul 8.6. Într-un experiment s-a urmărit efectul tipului de hrană și al stresului ambiental asupra greutății unor șobolani de laborator. Pentru aceasta s-au format 4 grupuri a câte 10 șobolani, care au fost supuse următoarelor tratamente: un grup a primit hrană normală în condiții nestresante, un grup a primit hrană normală în condiții stresante, un grup a primit hrană hipercalorică în condiții nestresante și ultimul grup a primit hrană hipercalorică în condiții stresante. După un timp s-au determinat greutățile la toți indivizii. Există o interacțiune semnificativă între cei doi factori: tipul de hrană și stres?

Având în vedere modul de organizare al datelor, se poate folosi ANOVA bifactorială cu replicare (cu număr egal de observații în celulă). La intersecția unui rând și a unei coloane sunt zece valori. Acest model poate răspunde la întrebarea problemei.

	Hrană	
Mediu	Normală	Hipercalorică
Stresant	130	157
	142	155
	131	162
	124	153
	124	158
	131	152
	143	159
	131	161
	130	160
	127	157
Nestresant	120	132
	130	128
	122	142
	118	131
	120	135
	119	120
	127	139
	118	133
	119	135
	122	136

Pentru a vedea dacă există interacțiune între factori la nivel de probe (celule), trebuie reprezentate grafic mediile valorilor din fiecare celulă.



Din analiza acestui grafic rezultă că hrana hipercalorică determină o creștere în greutate mai mare decât cea normală. Condițiile stresante determină o creștere în greutate mai mare. Creșterea în greutate când s-a administrat hrană hipercalorică și în condiții stresante este mai puternică decât atunci când s-a administrat hrană normală în condiții stresante. Deci s-ar putea să existe o interacțiune între cei doi factori.

Ipotezele acestui test sunt:

H_{0_1}: tipul de hrană nu are un efect semnificativ asupra greutății (coloanele nu diferă semnificativ);

H_{0_2}: stresul nu are un efect semnificativ asupra greutății (rândurile nu diferă semnificativ);

H_{0_3}: interacțiunea dintre tipul de hrană și stres nu este semnificativă (celulele nu diferă semnificativ);

H_{1_1}: tipul de hrană are un efect semnificativ asupra greutății (coloanele diferă semnificativ);

H_{1_2}: stresul are un efect semnificativ asupra greutății (rândurile diferă semnificativ);

H_{1_3}: interacțiunea dintre tipul de hrană și stres este semnificativă (celulele diferă semnificativ).

Pentru a verifica dacă interacțiunea este semnificativă trebuie să se parcurgă etapele de realizare ale ANOVA bifactorială cu replicare.

Se calculează suma pătratelor tuturor valorilor:

$$\sum x_t^2 = 320544 + 425335 = 745879 .$$

Se calculează pătratul sumei totale:

$$(\sum x_t)^2 = (2528 + 2905)^2 = (5433)^2 = 29517489 .$$

Se calculează numărul total de valori din toate probele:

$$n_t = 20 + 20 = 40 .$$

Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărei coloane:

$$\sum\left[\frac{\left(\sum x_{c_i}\right)^2}{n_{c_i}}\right]=\frac{(2528)^2}{20}+\frac{(2905)^2}{20}=741490{,}45\,.$$

Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărui rând:

$$\sum\left[\frac{\left(\sum x_{r_j}\right)^2}{n_{r_j}}\right]=\frac{(2887)^2}{20}+\frac{(2546)^2}{20}=740844{,}25\,.$$

Se calculează suma rapoartelor dintre pătratul sumei și dimensiunea fiecărei celule:

$$\sum\left[\frac{\left(\sum x_{c_i r_j}\right)^2}{n_{c_i r_j}}\right]=\frac{(1313)^2}{10}+\frac{(1574)^2}{10}+\frac{(1215)^2}{10}+\frac{(1331)^2}{10}=744923{,}1\,.$$

Se calculează sumele de pătrate:

$$SP_t=745879-\frac{(5433)^2}{40}=7941{,}775$$
$$SP_c=741490{,}45-\frac{(5433)^2}{40}=3553{,}225$$
$$SP_r=740844{,}25-\frac{(5433)^2}{40}=2907{,}025$$
$$SP_i=744923{,}1-\frac{(5433)^2}{40}-(3553{,}225+2907{,}025)=525{,}625$$
$$SP_{int}=7941{,}775-(3553{,}225+2907{,}025+525{,}625)=955{,}9.$$

Se calculează numărul gradelor de libertate:

$$gl_t=40-1=39 \qquad gl_c=2-1=1$$
$$gl_r=2-1=1 \qquad gl_i=(2-1)(2-1)=1$$
$$gl_{int}=40-4=36\,.$$

Se calculează sumele de pătrate medii:

$$\overline{SP_c} = \frac{3553,225}{1} = 3553,225 \qquad \overline{SP_r} = \frac{2907,025}{1} = 2907,025$$
$$\overline{SP_i} = \frac{525,625}{1} = 525,625 \qquad \overline{SP_{int}} = \frac{955,9}{36} = 26,553 .$$

Se calculează statisticile testului și se completează tabelul ANOVA:

$$F_c = \frac{3553,225}{26,553} = 133,82 \qquad F_r = \frac{2907,025}{26,553} = 109,48$$
$$F_i = \frac{525,625}{26,553} = 19,80 .$$

Sursa de variație	SP	gl	$\overline{SP}$	F
Externă, între coloane	3553,225	1	3553,225	133,82
Externă, între rânduri	2907,025	1	2907,025	109,48
Externă, de interacțiune (între celule)	525,625	1	525,625	19,80
Internă	955,9	36	26,553	
Totală	7941,775	39		

Deoarece gradele de libertate externe sunt toate egale cu 1, valoarea critică va fi aceeași (anexa 2 sau 3) pentru toate cele trei seturi de ipoteze: $F_{(0,05,1,36)} = 4,113$.

$133,82 > 4,113 \Rightarrow H_{0_1}$ se respinge și se acceptă H_{1_1} pentru o probabilitate $p = 1 - 0,05$.
$109,48 > 4,113 \Rightarrow H_{0_2}$ se respinge și se acceptă H_{1_2} pentru o probabilitate $p = 1 - \alpha$.
$19,80 > 4,113 \Rightarrow H_{0_3}$ se respinge și se acceptă H_{1_3} pentru o probabilitate $p = 1 - \alpha$.

Probabilitățile ca ipotezele nule să fie adevărate pot fi calculate pentru fiecare statistică F (anexa 3). Ele sunt $1,1 \cdot 10^{-13}$ pentru F_c, $1,9 \cdot 10^{-12}$ pentru F_r și $7,9 \cdot 10^{-5}$ pentru F_i.

Concluzia testului este că hrana are un efect semnificativ asupra

greutății, stresul are un efect semnificativ asupra greutății și interacțiunea dintre hrană și stres este și ea semnificativă.

În continuare, pentru evidențierea diferențelor semnificative dintre celule luate câte două se poate realiza testul Tukey. Pentru aceasta se calculează modulele diferențelor dintre mediile celulelor:

Hrană Mediu	Hipercalorică Stresant 157,4	Normală Nestresant 121,5	Hipercalorică Nestresant 133,1
Normală Stresant 131,3	26,1	9,8	1,8
157,4		35,9	24,3
121,5			11,6

Se află valoarea critică Tukey (anexa 2):

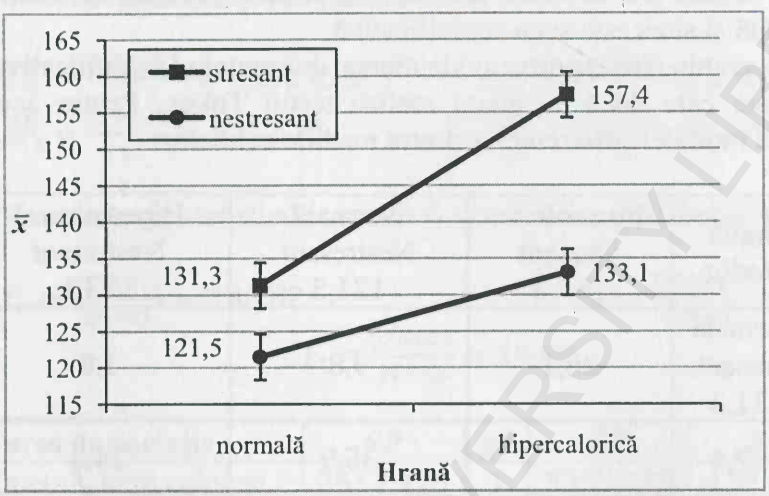
$q_{(0,05,4,36)} = 3,85$.

Se calculează statistica testului:

$$T = 3,85 \cdot \sqrt{\frac{26,55}{10}} = 6,27 .$$

Comparația dintre modulele diferențelor dintre medii și statistica testului evidențiază că există diferențe semnificative între toate perechile de grupuri, cu excepția perechii formate de grupul hrănit normal în condiții de stres și grupul hrănit cu hrană hipercalorică în condiții nestresante, între care nu există o diferență semnificativă.

Diferențele semnificative pot fi observate și prin analiza grafică a suprapunerii dintre intervalele de confidență ale mediilor grupurilor. Limitele intervalelor de confidență se calculează scăzând și adunând la media fiecărui grup valoarea obținută prin împărțirea statisticii testului Tukey la doi ($\bar{x}_{c_i r_j} \pm \frac{6,27}{2} = \bar{x}_{c_i r_j} \pm 3,135$).



9. CORELAȚIA ȘI REGRESIA

O cercetare ecologică poate urmări posibilele relații între două sau mai multe fenomene. De cele mai multe ori se urmărește relația dintre două variabile apreciate pe o scală ordinală, de interval sau de raport. Analiza unei astfel de relații se face prin corelație sau regresie, aplicarea uneia din cele două depinzând de modalitatea de obținere a datelor și de problema care se pune în legătură cu acestea.

Corelația este folosită pentru a determina dacă există asociere între două variabile și cât de puternică este această asociere. Prin asociere se înțelege că atunci când o variabilă se modifică cealaltă se modifică și ea într-un anumit mod. De remarcat că în cazul corelației nu se fac presupuneri vizând asocieri de tipul cauză-efect între cele două variabile, deși acestea ar putea exista. Există posibilitatea ca dinamica celor două variabile să fie determinată de o a treia.

Regresia, pe de altă parte, evidențiază cu precădere relațiile de tip cauză-efect dintre două variabile, astfel încât o proporție substanțială dintre valorile unei variabile, numită **variabilă dependentă**, să fie o funcție sau să fie explicate de valorile celeilalte variabile, numită **variabilă independentă**.

O altă deosebire notabilă dintre corelație și regresie este faptul că în majoritatea cazurilor de analiză a regresiei valorile variabilei independente nu sunt obținute aleator din populație, nu sunt normal distribuite, ci, mai curând, selecția lor se află sub controlul experimentatorului.

În general, se fac suficiente confuzii privind care dintre cele două analize pot fi aplicate unor anumite date. Pentru a simplifica decizia privind utilizarea analizei corelației sau regresiei să analizăm următoarele trei cazuri:

A. Se extrage **aleator** o probă formată din șopârle gravide dintr-o populație. După depunerea ouălor se înregistrează greutatea și numărul de ouă produse. Datele se reprezintă grafic, desemnându-se **arbitrar** care dintre cele două variabile va fi reprezentată pe abscisă și care pe ordonată.

B. Se urmărește efectul temperaturii asupra frecvenței cardiace la

proba reprezentată de șopârle. Pentru aceasta fiecare dintre ele este supusă unei **anumite** temperaturi cuprinsă între anumite limite. Se înregistrează apoi la fiecare individ frecvența cardiacă. Temperatura, **variabila independentă fixată arbitrar**, va fi reprezentată pe **abscisă**. Frecvența cardiacă, **variabila dependentă** de prima, va fi reprezentată pe **ordonată**.

C. Se alege un anumit număr de șopârle gravide dintr-o populație **după un anumit criteriu** – să aibă o anumită greutate. Se urmărește mai departe câte ouă va produce fiecare animal. Greutatea, **variabila independentă aleasă arbitrar** și în funcție de valorile căreia s-a selecționat proba, va fi reprezentată pe **abscisă**, iar dimensiunea pontei, **variabila dependentă** de prima, va fi reprezentată pe **ordonată**.

Dintre cele 3 situații descrise mai sus, prima (**A**) este o problemă de analiză a corelației, în timp ce situațiile **B** și **C** sunt probleme de analiză a regresiei. Deși situațiile **A** și **C** pot conduce aparent la ideea că ambele analize pot fi aplicate pe aceleași date, în realitate, aplicarea corelației sau regresiei este dictată de modul în care a fost obținută proba: în cazul **A** proba era prelevată aleator, în timp ce în cazul **C** proba era prelevată arbitrar, după un anumit criteriu.

9.1. ANALIZA CORELAȚIEI

Analiza corelație are rolul de a răspunde la mai multe întrebări:

1. Există o relație între două 2 variabile studiate?
2. Care este tipul acestei relații?
3. Cât de puternică este această relație?
4. Este relația detectată semnificativă?

Răspunsurile la aceste întrebări pot fi intuite din analiza datelor prin intermediul unor reprezentări grafice (diagramă de împrăștiere a punctelor de coordonate x și y). Astfel, analiza graficelor de mai jos (fig. 9.1) permite următoarele concluzii:

– primul grafic prezintă o corelație directă sau pozitivă (variabilele sunt direct proporționale), adică cele două variabile se modifică în același sens – când x crește, y crește, iar când x scade, y scade;

– al doilea grafic prezintă o corelație inversă sau negativă (variabilele sunt invers proporționale), adică cele două variabile se modifică în sens diferit – când x crește, y scade, iar când x scade, y crește;

– al treilea grafic arată că, între cele două variabile, corelația practic nu există;

– în cazul celui de-al patrulea grafic, tragerea unei concluzii este dificilă, implicând un grad înalt de subiectivism: fie nu există corelație, fie este o corelație pozitivă, slabă (împrăștierea punctelor este relativ mare).

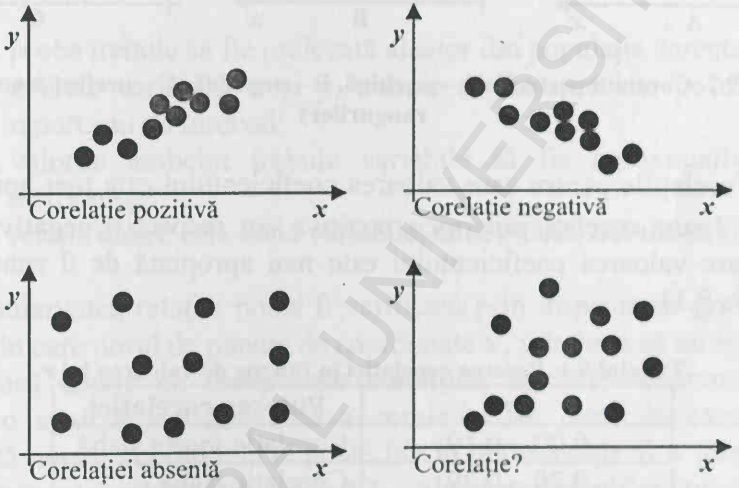


Figura 9.1. Tipuri de corelație

Examinarea graficelor de corelație poate fi însă subiectivă, motiv pentru care este necesară completarea metodei grafice cu una statistică, mai obiectivă. O astfel de metodă ce surprinde măsura în care două variabile sunt asociate constă în calcularea **coeficientului de corelație** r.

Valoarea coeficientului de corelație este cuprinsă între -1, valoarea unei corelații maxime negative, și $+1$, valoarea unei corelații maxime pozitive. Dacă r este cuprins între 0 și $+1$, corelația este pozitivă, iar dacă este cuprins între 0 și -1, corelația este negativă. Când r este egal cu 0, corelația este absentă. Când r este egal cu $+1$ sau -1, punctele de coordonate x, y sunt dispuse perfect liniar, în lungul unei drepte imaginare (fig. 9.2, A și B). Dacă variabilele urmărite sunt apreciate pe o scală

ordinală, există o corelație perfectă când toate valorile cresc sau descresc succesiv. În acest caz, punctele din graficul de corelație pot să nu se afle dispuse liniar (fig. 9.2, C).

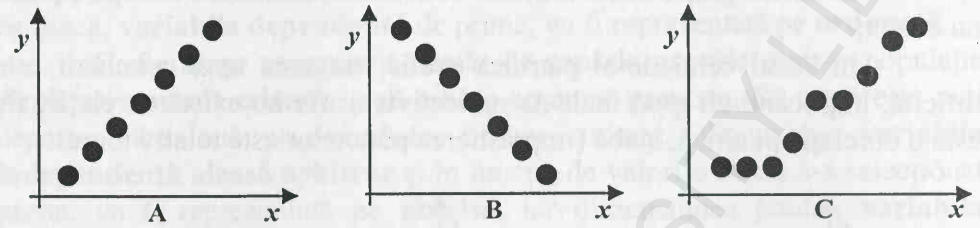


Figura 9.2. Corelații perfecte: A – pozitivă, B – negativă, C – perfect monotonă (a rangurilor)

Corelațiile pentru care valoarea coeficientului este mai apropiată de +1 sau – 1 sunt corelații puternice, pozitive sau, respectiv, negative, iar cele pentru care valoarea coeficientului este mai apropiată de 0 sunt corelații slabe (tab. 9.1).

Tabelul 9.1. Puterea corelației în funcție de valoarea lui *r*

$\pm r$	**Puterea corelației**
0,00 – 0,19	Corelație foarte slabă
0,20 – 0,39	Corelație slabă
0,40 – 0,69	Corelație moderată
0,70 – 0,89	Corelație puternică
0,90 – 1,00	Corelație foarte puternică

În funcție de caracteristicile datelor celor două variabile, corelația poate fi parametrică sau neparametrică. Pentru fiecare din cele două se folosesc coeficienți diferiți: pentru corelația parametrică se folosește **coeficientul de corelație Pearson**; pentru corelația neparametrică se utilizează **coeficientul de corelație Spearman**. În funcție de utilizarea unuia dintre cei doi coeficienți, analiza corelației poate fi **parametrică** și, respectiv, **neparametrică**. Analiza corelație în general constă în calcularea valorii coeficientului de corelație și în testarea semnificației acestuia la nivelul populației din care a fost extrasă proba.

Semnificația unei corelații se apreciază independent de puterea acesteia. Astfel, se poate pune în evidență o corelație puternică, pozitivă sau negativă, care însă să nu fie semnificativă și invers, una slabă care însă să se dovedească a fi semnificativă.

9.1.1. Analiza corelației parametrice

Ca și în cazul celorlalte teste parametrice prezentate în capitolele anterioare, analiza corelației parametrice presupune ca datele să îndeplinească următoarele condiții:

1. proba trebuie să fie prelevată aleator din populația cercetată;
2. ambele variabile, x și y, trebuie să fie apreciate pe o scală de raport sau de interval;
3. valorile ambelor trebuie variabile să fie aproximativ normal distribuite;
4. relația dintre cele două variabile, dacă există, trebuie să fie liniară.

Liniaritatea relației poate fi verificată prin inspectarea graficului de corelație în care norul de puncte de coordonate x, y trebuie să nu fie curbat.

Dacă datele nu îndeplinesc condițiile 2, 3, și 4, atunci trebuie utilizată o analiză neparametrică a corelației. În cazul în care nu este îndeplinită numai condiția 4, se poate lua în considerație și o transformare care să „îndrepte" relația (secțiunea 9.2 – Abordarea relațiilor curbilinii).

În cazul corelației parametrice se calculează coeficientul de corelație Pearson. Acesta măsoară cât de puternică este relația dintre două variabile x și y, pornind de la **covarianța** lor în probă.

Covarianța este o măsură a variabilității legate a două variabile față de mediile lor. Este de fapt media produselor abaterilor celor două variabile față de media fiecăreia sau produsul deviațiilor variabilelor. Ca și deviația standard și varianța, covarianța poate fi o statistică a unei probe (s_{xy}) sau un parametru populațional (σ_{xy}). Când statistica se folosește ca estimator al parametrului, atunci suma produselor abaterilor variabilelor se împarte la numărul gradelor de libertate $n-1$.

$$\sigma_{xy} = \frac{\sum(x-\mu_x)(y-\mu_y)}{n} \qquad s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

Pentru două valori x și y, fiecare mai mare decât media sa ($x > \bar{x}$ și $y > \bar{y}$), abaterile ($x - \bar{x}$ și $y - \bar{y}$) lor vor fi pozitive și produsul lor, tot pozitiv. Dacă o pereche de valori x și y sunt mai mici decât mediile lor, atunci ambele abateri vor fi negative. Produsul lor însă va fi pozitiv.

Dacă o valoare x este mai mică decât $\bar{x}$, iar o valoare y este mai mare decât $\bar{y}$, atunci abaterea lui x va fi negativă, iar abaterea lui y va fi pozitivă. Ca urmare, produsul abaterilor va fi negativ. La fel se întâmplă dacă x este mai mare decât $\bar{x}$, iar y este mai mic decât $\bar{y}$.

Din punct de vedere grafic, produsele abaterilor vor fi pozitive sau negative în funcție de poziția punctului de coordonate x, y față de poziția punctului de coordonate $\bar{x}$, $\bar{y}$, numit și **centru mediu** (fig. 9.3).

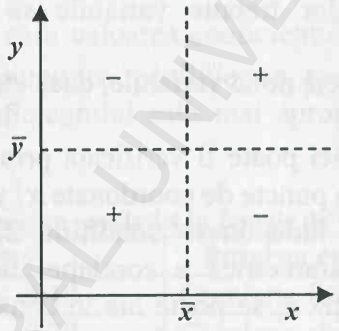


Figura 9.3. Semnul produselor abaterilor în funcție de poziția față de centrul mediu

Referindu-ne la datele dintr-o probă, din analiza figurii 9.3 rezultă că, dacă majoritatea punctelor de coordonate x, y se vor dispune în cadranele notate cu „+" și mai puține puncte se vor afla în cadranele notate cu „–", atunci covarianța va fi pozitivă. Dacă vor predomina punctele din cadranele negative, iar cele din cadranele pozitive vor fi mai puține, atunci covarianța probei va fi negativă. Deci, făcând legătura dintre covarianță și coeficientul de corelație, se poate spune că semnul corelației, al coeficientului, este același cu cel al covarianței valorilor dintr-o probă. Acest lucru este evident și dacă suprapunem primele două grafice din figura 9.1 peste cel din figura 9.3 (fig. 9.4).

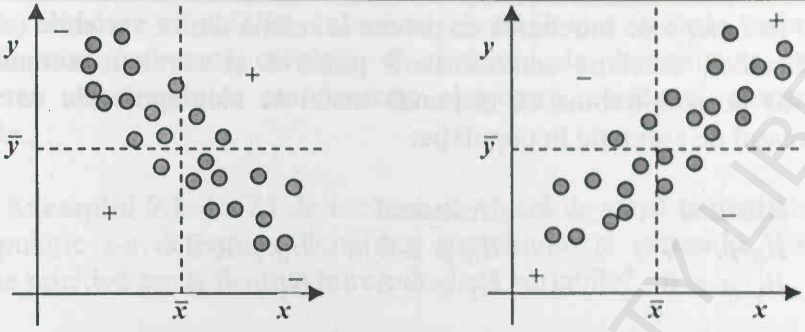


Figura 9.4. Relația dintre semnul covarianței și cel al corelației

Se poate pune următoarea întrebare: dacă relația dintre variabile este surprinsă de covarianța lor, de ce nu se utilizează acest descriptor (s_{xy}) în loc de coeficientul de corelație (r)? Din perspectiva corelației, covarianța prezintă un dezavantaj: valoarea sa este influențată de unitățile de măsură ale variabilelor x și y, ceea ce face dificilă comparația. Pentru a elimina acest neajuns, este nevoie să se realizeze o standardizare a covarianței (s_{xy}) prin împărțirea acesteia la produsul deviațiilor standard ale celor două variabile ($s_x \cdot s_y$). Astfel, câtul acestei împărțiri va lua valori de la − 1 la +1 și va reprezenta **coeficientul de corelație Pearson**:

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}}{\sqrt{\frac{\sum(x-\bar{x})^2}{n-1} \cdot \frac{\sum(y-\bar{y})^2}{n-1}}}.$$

Prin rearanjarea algebrică a formulei se obține una mai ușor de folosit în practică:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}.$$

Odată calculată valoarea coeficientului de corelație se poate trece la testarea semnificației acestuia care se face sub forma unui test. Testarea semnificației arată în ce măsură relația surprinsă de coeficientul de corelație a probei (r) este semnificativă la nivel de populație (ρ).

În funcție de întrebarea cu privire la relația dintre variabile (corelație semnificativă, corelație semnificativă pozitivă și corelație semnificativă negativă) la care trebuie să răspundă testul se aleg ipotezele ce vizează coeficientul de corelație în populație:

Bilateral	Unilateral	
$H_0: \rho = 0$	$H_0: \rho \leq 0$	$H_0: \rho \geq 0$
$H_1: \rho \neq 0$	$H_1: \rho > 0$	$H_1: \rho < 0$.

Primul set de ipoteze se folosește dacă se dorește evidențierea unei corelații semnificative în populația din care s-a extras proba, fără a se preciza semnul acesteia. Dacă se urmărește punerea în evidență a unei corelații pozitive semnificative, se utilizează al doilea set de ipoteze, iar în cazul în care se urmărește semnificația unei corelații negative, se folosește ultimul set de ipoteze.

Statistica testului este una de tip Student (t) și se calculează cu ajutorul numărului de perechi de valori x și y (n) și cu valoarea coeficientului de corelație Pearson (r):

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

Condiția testului este similară cu cea a unui test t, cu mențiunea că valoarea critică se alege în funcție de α și de numărul gradelor de libertate $n-2$.

Dacă $|t| \geq t_{(\alpha,n-2)} \Rightarrow H_0$ se respinge, H_1 se acceptă pentru $p = 1-\alpha$.

Coeficientul de determinare r^2

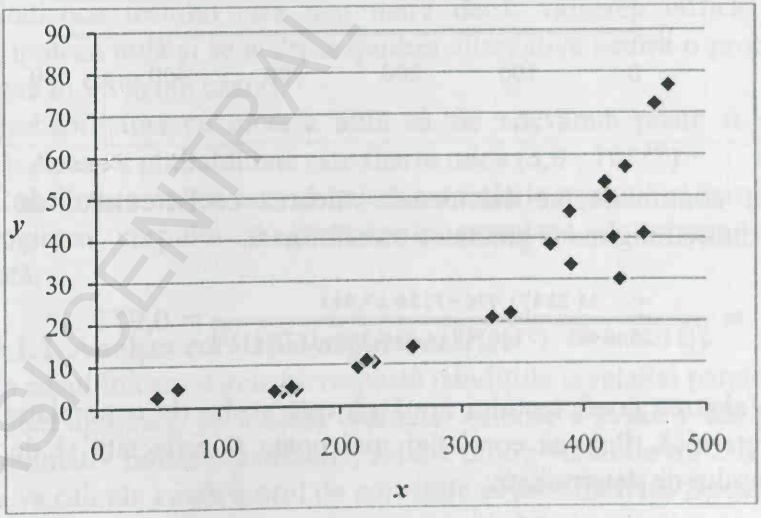
Valoarea coeficientului de corelație Pearson ridicată la pătrat reprezintă o statistică utilă a datelor. Acesta arată proporția în care variabilitatea uneia dintre cele două variabile poate fi pusă pe seama variabilității celeilalte. Coeficientul de determinare reprezintă o proporție, dar dacă se înmulțește cu 100, rezultă procentul de valori ale celor două variabile care sunt realmente corelate. De exemplu, dacă pentru două

variabile oarecare $r^2 = 0{,}81$, înseamnă că 81% din valorile celor două variabile sunt realmente corelate. Coeficientul de determinare poate fi considerat un descriptor standardizat al puterii corelației dintre două variabile.

Exemplul 9.1. La 24 de indivizi de viperă de stepă extrași aleatoriu din populație s-a determinat lungimea trunchiului și greutatea. Există o corelație pozitivă semnificativă între cele două variabile?

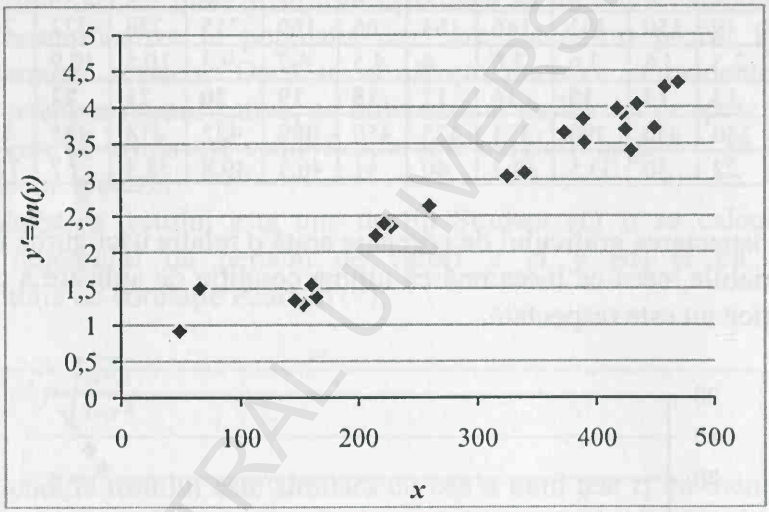
Nr. crt.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
cm (*x*)	49	150	153	146	164	66	160	215	228	222	260	325
g (*y*)	2,5	3,6	3,6	3,8	4	4,5	4,7	9,3	10,5	10,9	14	21
Nr. crt.	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
cm (*x*)	340	430	390	373	425	450	389	422	418	435	459	470
g (*y*)	22	30	33,5	38,4	40	41	46,3	49,8	53,3	57	72,2	76,7

Inspectarea graficului de corelație arată o relație ușor curbă între cele două variabile, ceea ce înseamnă că ultima condiție de aplicare a corelației parametrice nu este respectată.



În urma logaritmării valorilor x și y în diferite combinații se observă că logaritmarea doar a valorilor greutăților îndreaptă cel mai bine relația dintre cele două variabile.

Nr. crt.	1	2	3	4	5	6	7	8	9	10	11	12
x	49	150	153	146	164	66	160	215	228	222	260	325
$y'=ln(y)$	0,916	1,281	1,281	1,335	1,386	1,504	1,548	2,230	2,351	2,389	2,639	3,045
Nr. crt.	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
x	340	430	390	373	425	450	389	422	418	435	459	470
$y'=ln(y)$	3,091	3,401	3,512	3,648	3,689	3,714	3,835	3,908	3,976	4,043	4,279	4,340



În continuare, se calculează valoarea coeficientului de corelație Pearson dintre lungime și greutatea transformată:

$$r = \frac{24 \cdot 23471{,}976 - 7139 \cdot 67{,}341}{\sqrt{[24 \cdot 2546985 - (7139)^2][24 \cdot 218{,}368 - (67{,}341)^2]}} = 0{,}975 .$$

Valoarea coeficientului arată că este vorba de o corelație pozitivă foarte puternică. Puterea corelației mai poate fi reflectată și de valoarea coeficientului de determinare:

$$r^2 = 0{,}975^2 = 0{,}95 .$$

Valoarea coeficientului de determinare arată că 95% din valorile celor două variabile sunt realmente corelate.

Pentru a afla dacă această corelație evidențiată la nivel de probă este semnificativă și la nivel de populație, trebuie efectuat testul de semnificație a corelației. Având în vedere modul de formulare a întrebării problemei, ipotezele testului sunt:

$H_0: \rho \leq 0$
$H_1: \rho > 0$.

Se calculează statistica testului:

$$t = 0{,}975 \cdot \sqrt{\frac{24-2}{1-0{,}95}} = 20{,}54 .$$

Se află valoarea critică pentru un test unilateral (anexa 2 sau 3):

$$t_{(0{,}05,24-2)} = 1{,}717 .$$

Statistica testului este mai mare decât valoarea critică, deci se respinge ipoteza nulă și se acceptă ipoteza alternativă pentru o probabilitate de 0,95 sau în 95% din cazuri.

Probabilitatea ca ipoteza nulă să fie adevărată poate fi calculată (anexa 3). Această probabilitate este foarte mică ($3{,}8 \cdot 10^{-16}$).

Concluzia testului este că există o corelație pozitivă și semnificativă între lungimea corpului și greutate în populația de vipere de stepă investigată.

9.1.2. Analiza corelației neparametrice

În cazul în care datele nu respectă condițiile corelației parametrice (x și/sau y sunt apreciate pe o scală ordinală, valorile x și/sau y din probă nu sunt aproximativ normal distribuite, relația dintre variabile nu este liniară), atunci se va calcula **coeficientul de corelație neparametrică Spearman**.

Acest coeficient, notat cu r_s pentru probă și ρ pentru populație, a fost obținut prin prelucrarea formulei coeficientului Pearson când în loc de

datele brute se folosesc rangurile acestora. Deci există o legătură strânsă între cei doi coeficienți.

Pentru a putea calcula valoarea r_s este nevoie ca mai întâi să se dea ranguri valorilor lui x (R_x) și valorilor lui y (R_y), separat. Algoritmul de acordare a rangurilor este similar cu cel folosit în celelalte teste neparametrice (secțiunea 2.1, tab. 2.3). Ulterior, se calculează diferențele dintre rangurile corespunzătoare fiecărei perechi de valori x și y, originale $(d = R_x - R_y)$, iar fiecare diferență se ridică la puterea a doua (d^2). Pătratele tuturor diferențelor se însumează obținându-se suma pătratelor diferențelor $(\sum d^2)$. Formula de calcul a coeficientului Spearman include suma pătratelor diferențelor și numărul perechilor de valori x și y (n):

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}.$$

Valoarea coeficientului Spearman are aceleași proprietăți informaționale ca și cea a coeficientului Pearson: ia valori între -1 și $+1$, semnul valorii indică, după caz, o corelație pozitivă sau negativă; o valoare egală cu $+1$ sau -1 semnifică o corelație maximă, în timp ce una egală cu 0 arată absența corelației; apropierea valorii de ± 1 indică o corelație puternică, în timp ce o valoare mai apropiată de 0, o corelație slabă.

Testarea semnificației în analiza corelației neparametrice se face la fel ca în cazul corelației parametrice, cu deosebirea că la calcularea statisticii testului (t) în loc de valoarea coeficientului Pearson (r) se folosește valoarea coeficientului Spearman (r_s):

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}.$$

Coeficientul Spearman pierde din precizie dacă există un număr relativ mare de valori egale. În general, dacă mai mult de jumătate din ranguri au valori egale, atunci se recomandă calcularea coeficientului de corelație a rangurilor conform formulei lui Pearson, în care se înlocuiesc valorile originale cu rangurile acestora.

Exemplul 9.2. Să se rezolve problema de la exemplul 9.1 considerându-se că datele nu îndeplinesc condițiile de aplicare ale corelației parametrice.

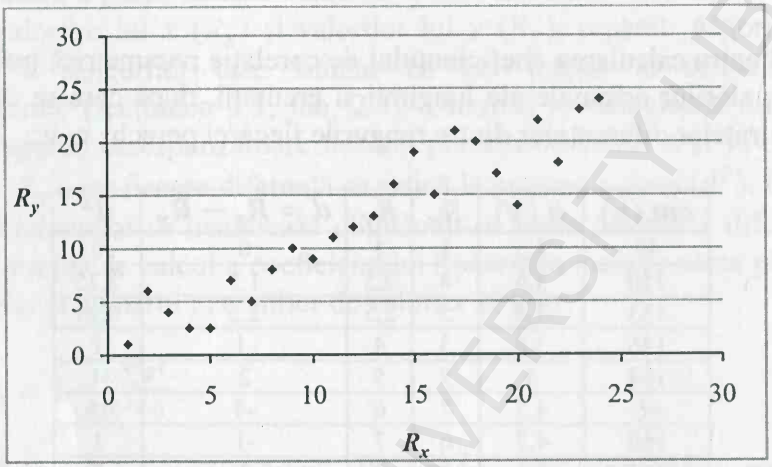
Pentru calcularea coeficientului de corelație parametrică trebuie date ranguri valorilor originale ale lungimii și greutății, după care se calculează suma pătratelor diferențelor dintre rangurile fiecărei perechi x, y:

$cm\ (x)$	$g\ (y)$	R_x	R_y	$d = R_x - R_y$	d^2
49	2,5	1	1	0	0
150	3,6	4	2,5	1.5	2,25
153	3,6	5	2,5	2.5	6,25
146	3,8	3	4	-1	1
164	4	7	5	2	4
66	4,5	2	6	-4	16
160	4,7	6	7	-1	1
215	9,3	8	8	0	0
228	10,5	10	9	1	1
222	10,9	9	10	-1	1
260	14	11	11	0	0
325	21	12	12	0	0
340	22	13	13	0	0
430	30	20	14	6	36
390	33,5	16	15	1	1
373	38,4	14	16	-2	4
425	40	19	17	2	4
450	41	22	18	4	16
389	46,3	15	19	-4	16
422	49,8	18	20	-2	4
418	53,3	17	21	-4	16
435	57	21	22	-1	1
459	72,2	23	23	0	0
470	76,7	24	24	0	0
				$\sum d^2 =$	130,5

Se calculează coeficientul de corelație Spearman:

$$r_s = 1 - \frac{6 \cdot 130,5}{24^3 - 24} = 0,943\ .$$

Acest rezultat arată că este vorba de o corelație pozitivă foarte puternică.



Pentru testarea semnificației se aplică testul la fel ca în exemplul 9.1.
Se scriu ipotezele testului:
$H_0: \rho \leq 0$
$H_1: \rho > 0$.

Se calculează statistica testului:

$$t = 0{,}943 \cdot \sqrt{\frac{24-2}{1-(0{,}943)^2}} = 13{,}324 .$$

Se află valoarea critică pentru testul unilateral (anexa 2 sau 3):

$$t_{(0{,}05,24-2)} = 1{,}717 .$$

$13{,}324 > 1{,}717 \Rightarrow H_0$ se respinge și se acceptă H_1 pentru o probabilitate de 0,95.

Probabilitatea ipotezei nule (anexa 3) este de $2{,}6 \cdot 10^{-12}$.

Concluzia testului este că există o corelație pozitivă semnificativă între lungimea și greutatea viperelor de stepă din populația studiată.

9.2. ANALIZA REGRESIEI

Analiza regresiei este similară în unele aspecte cu analiza corelației, dar este diferită de aceasta prin faptul că presupune o relație de tip cauză-efect între variabila independentă, aflată sub controlul cercetătorului, și cea dependentă.

În esență, se presupune că ar exista o relație funcțională care permite prezicerea unei valori a variabilei y corespunzătoare unei valori date a variabilei independente x, adică:

$y = f(x) + e$ e – eroarea aleatoare.

În cazul regresiei liniare simple, relația are următoarea formă:

$\mu_y = \alpha + \beta x$

μ_y – media populațională a valorilor y corespunzătoare unei valori x

α – coeficient de regresie – înălțimea dreptei de regresie

β – coeficient de regresie – panta dreptei de regresie.

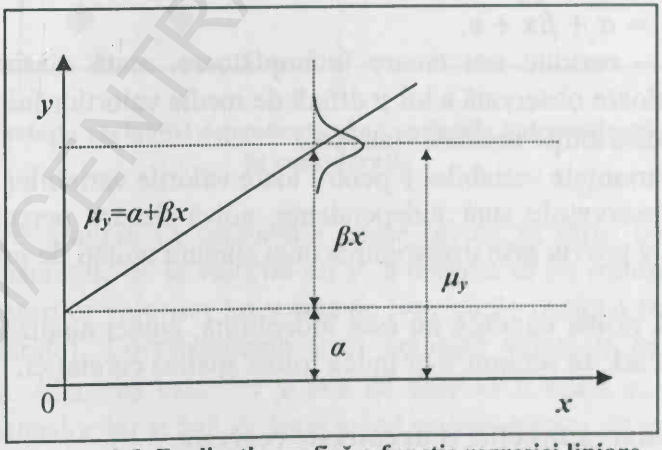


Figura 9.5. Explicația grafică a funcție regresiei liniare

Analiza regresiei presupune următoarele aspecte:

1. Regresia este folosită pentru aproximarea unei ecuații ce descrie relația liniară dintre 2 variabile. Aceasta se numește **ecuație** sau **funcție de regresie**. Parametrii α și β pot fi estimați pe baza unei probe din populație.
2. Pe baza ecuației se poate construi o dreaptă de regresie.
3. Ecuația de regresie poate fi folosită pentru aflarea valorilor variabilei dependente (y) corespunzătoare valorilor variabilei independente (x).
4. Regresia poate fi folosită pentru a surprinde măsura în care variabila dependentă se află sub controlul variabilei independente.

Analiza regresiei presupune o serie de condiții:

1. Variabila independentă (x) este fixată, adică valorile acesteia sunt alese arbitrar și nu aleator din populație.
2. Pentru orice valoare a variabilei independente (x) există o populație normal distribuită de valori ale variabilei dependente (y). Media populațională a valorilor lui y este: $\mu_y = \alpha + \beta x$.
3. Din condiția 2 rezultă că pentru oricare valoare x există o valoare particulară y_i:
4. $y_i = \alpha + \beta x + e$.
5. e – reziduu sau eroare întâmplătoare; arată măsura în care o valoare observată a lui y diferă de media valorilor lui y (μ_y); e are o distribuție normală standard.
6. Variațele variabilei y pentru toate valorile variabilei x sunt egale.
7. Observațiile sunt independente, adică fiecare pereche de valori x, y provin prin investigarea unei singure unități de probă.

Dacă prima condiție nu este îndeplinită, atunci analiza regresiei nu poate fi aplicată. În schimb, s-ar putea folosi analiza corelației.

Estimarea funcției și dreptei de regresie

Coeficienții de regresie, adică parametrii funcției, se estimează pornind de la o probă, astfel α și β vor fi estimați prin a și, respectiv, b.

Dreapta descrisă de estimatorii a și b reprezintă linia care se potrivește cel mai bine funcției de regresie și care se mai numește și **dreaptă de regresie estimată**. Având în vedere că ecuația definește dreapta, estimarea acestora se va face în paralel.

Dreapta de regresie va trece întotdeauna prin punctul de coordonate $\bar{x}, \bar{y}$. Dacă pe orizontala ce trece prin acest punct se trasează perpendiculare din fiecare punct de coordonate x, y, atunci fiecare perpendiculară va reprezenta abaterea valorilor lui y față de $\bar{y}$, adică $y-\bar{y}$. Suma acestor abateri va fi aproximativ egală cu 0, dar suma pătratelor abaterilor va fi însă mai mare decât 0 ($\sum(y-\bar{y}) \approx 0$; $\sum(y-\bar{y})^2 > 0$). Deci se obține suma pătratelor abaterilor lui y fără a se ține cont de valorile lui x (fig. 9.6).

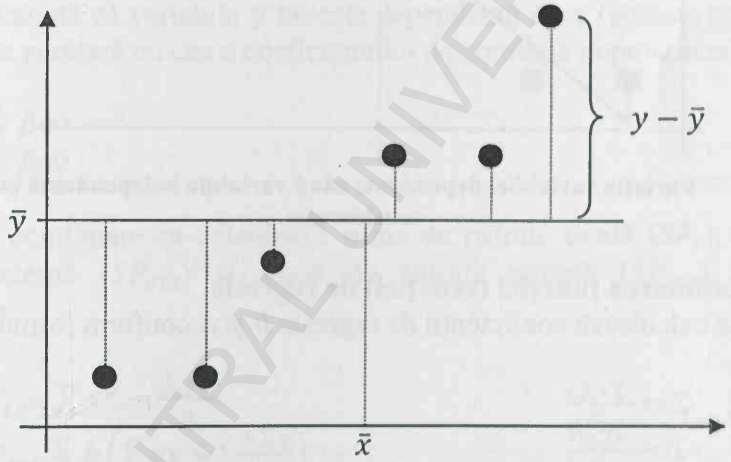


Figura 9.6. Variația variabilei dependente când variabila independentă nu este luată în considerație

Să presupunem că orizontala poate pivota în jurul punctului $\bar{x}, \bar{y}$ astfel încât distanțele de la valorile lui y la dreaptă să fie minime sau suma pătratelor abaterilor valorilor lui y față de linie să fie minimă (metoda celor mai mici pătrate). Corespondenta unei anumite valori y pe dreaptă se notează cu $\hat{y}$. Abaterea valorilor y față de linie va fi egală cu $y-\hat{y}$. Suma pătratelor abaterilor lui y față de linie, când se ține cont și de valorile lui x, va fi mai mică decât cea precedentă, adică $\sum(y-\hat{y})^2 < \sum(y-\bar{y})^2$. În concluzie, incertitudinea lui y a fost redusă prin luarea lui x în considerație.

Diferența dintre o valoare y și corespondenta sa pe dreapta de regresie este $y-\hat{y}=e$; e are o distribuție normală și media egală cu 0, adică are o distribuție normală standard (fig. 9.7).

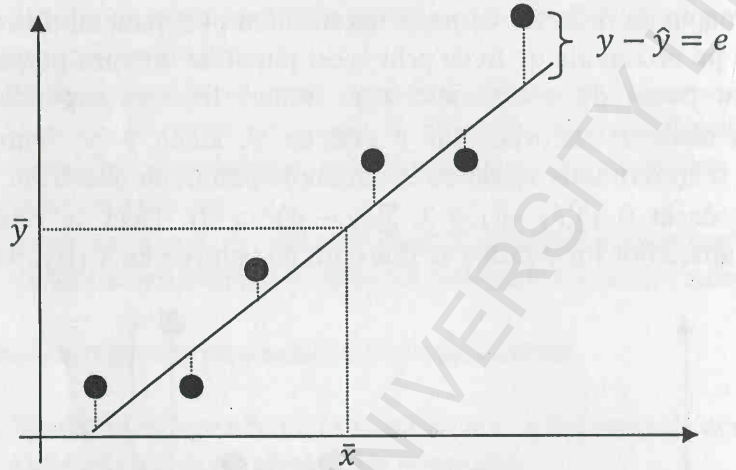


Figura 9.7. Variația variabilei dependente când variabila independentă este luată în considerație

Estimarea funcției (ecuației) de regresie

Se calculează coeficienții de regresie b și a conform formulelor:

$$b=\frac{\Sigma xy-\frac{\Sigma x\Sigma y}{n}}{\Sigma x^2-\frac{(\Sigma x)^2}{n}}$$

$$a=\bar{y}-b\bar{x}\,.$$

Estimarea funcției de regresie este:

$$\hat{y}=a+bx\,.$$

Cu ajutorul funcției de regresie se calculează valorile $\hat{y}$ corespunzătoare fiecărei valori x. Trasarea dreptei de regresie se face unind punctele de coordonate $x, \hat{y}$.

Testarea semnificației funcției de regresie

Semnificația ecuației de regresie estimate se face sub forma unei analize a variației (ANOVA).

Principiul de descompunere a variabilității este următorul: varianța totală (s_t^2) este dată de varianța externă sau de regresie (s_{ext}^2) plus varianța internă sau reziduală (s_{int}^2):

$$s_t^2 = s_{ext}^2 + s_{int}^2 .$$

Ipotezele testului sunt similare cu cele de la corelație, doar că în cazul regresiei se testează coeficientul de regresie β (panta sau înclinația dreptei) pentru populația din care a fost prelevată proba. Dacă β este zero, atunci înseamnă că variabila y nu este dependentă de x (această proprietate a lui β este similară cu cea a coeficientului de corelație populațional ρ).

H_0: β=0
H_1: $\beta \neq 0$

În continuare se calculează suma de pătrate totală (SP_t), suma de pătrate externă (SP_{ext}) și suma de pătrate internă (SP_{int}), conform formulelor:

$$SP_t = \sum y^2 - \frac{(\sum y)^2}{n}$$
$$SP_{ext} = b\left(\sum xy - \frac{\sum x \sum y}{n}\right)$$
$$SP_{int} = SP_t - SP_{ext} .$$

Se calculează sumele de pătrate medii ($\overline{SP}$) ca raportul dintre sumele de pătrate (SP) și gradele de libertate corespunzătoare (gl). Gradele de libertate externe reprezintă diferența dintre numărul variabilelor considerate (k) din care se scade 1. Gradele de libertate totale reprezintă numărul perechilor de valori x, y (n) minus 1. Gradele de libertate interne se află la fel ca și suma de pătrate internă, adică din numărul gradelor de libertate totale se scad cele externe ($n-1-k+1 = n-k$). Raportul dintre suma de pătrate medie externă și cea internă reprezintă statistica testului (F).

Sursa de variabilitate	SP	gl	$\overline{SP}$	F
Externă	SP_{ext}	$k-1$	$\overline{SP}_{ext}$	$\overline{SP}_{ext}/\overline{SP}_{int}$
Internă	SP_{int}	$n-k$	$\overline{SP}_{int}$	
Totală	SP_t	$n-1$		

Condiția testului constă în compararea statisticii testului (F) cu o valoare critică pentru un anumit nivel de încredere α și gradele de libertate externe ($k-1$) și interne ($n-k$).

Dacă $F \geq F_{(\alpha,k-1,n-k)} \Rightarrow H_0$ se respinge, H_1 se acceptă pentru $p = 1-\alpha$.

Intervalul de confidență al coeficientului de regresie β

Coeficientul de regresie în probă b este o estimare a coeficientului de regresie populațional β. Deci b și β sunt diferiți, dar se poate calcula intervalul de confidență pentru β cu ajutorul erorii standard a lui b (s_b):

$$s_b = \sqrt{\frac{\overline{SP}_{int}}{\sum x^2 - \frac{(\sum x)^2}{n}}}.$$

Limitele inferioară (LI) și cea superioară (LS) ale intervalului de confidență a lui β se vor afla prin calcularea relației:

$$\beta = b \pm s_b \cdot t_{(\alpha,n-2)}$$
$$LI = b - s_b \cdot t_{(\alpha,n-2)}$$
$$LS = b + s_b \cdot t_{(\alpha,n-2)}.$$

În concluzie, intervalul LI-LS include coeficientul de regresie populațional β, cu o probabilitate de $1-\alpha$ sau $100(1-\alpha)$.

Coeficientul de determinare r^2

Se știe că varianța lui y poate fi explicată în bună măsură de cunoașterea variabilei x, totuși o parte rămâne neexplicată. Este vorba de

varianța internă. Dacă valorile lui y ar fi complet dependente de valorile lui x, atunci erorile aleatoare (e) ar fi egale cu zero, adică toate punctele de coordonate x, y ar fi exact pe dreapta de regresie.

În cazul analizei regresiei, coeficientul de determinare (r^2) arată proporția variației lui y explicată prin dependența de x.

Formula de calcul a lui r^2 este la fel ca în cazul analizei corelației. Totuși, pentru a simplifica calculul și pentru a folosi valori deja calculate în etapele anterioare ale analizei regresiei, r^2 se poate calcula și după următoarea formulă:

$$r^2 = \frac{SP_{ext}}{SP_t}.$$

Ca și în cazul analizei corelației, valoarea coeficientului de determinare se poate înmulți cu 100 pentru a obține un procent, care, în analiza regresiei, arată cât la sută dintre valorile lui y sunt dependente sau determinate de valorile lui x. Diferența $100-r^2\%$ arată varianța individuală sau reziduală care nu poate fi explicitată de valorile lui x.

Zona de confidență a dreptei de regresie

Limitele de confidență ale dreptei de regresie se pot afla calculând eroarea standard pentru fiecare punct de coordonate $x, \hat{y}$ de pe dreaptă, pentru fiecare valoare a lui x din probă. Pentru fiecare valoare x se calculează eroarea standard $s_{\hat{y}}$:

$$s_{\hat{y}} = \sqrt{\overline{SP}_{int}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}\right]}.$$

Limitele inferioară (LI) și superioară (LS) ale intervalului de confidență a fiecărui $\hat{y}$ corespunzător unei valori x din probă se calculează pornind de la relația:

$$\mu_{\hat{y}} = \hat{y} \pm s_{\hat{y}} \cdot t_{(\alpha,n-2)}$$
$$LI = \hat{y} - s_{\hat{y}} \cdot t_{(\alpha,n-2)}$$
$$LS = \hat{y} + s_{\hat{y}} \cdot t_{(\alpha,n-2)}.$$

Concluzia va fi că intervalele *LI-LS* pentru fiecare valoare x includ mediile populațiilor de valori $\hat{y}$ ($\mu_{\hat{y}}$) cu o probabilitate de $1-\alpha$.

Unirea limitelor inferioare între ele și a celor superioare între ele, obținute pentru fiecare valoare x, duce la reprezentarea grafică a zonei de confidență a dreptei în ansamblu (fig. 9.8). Deci dreapta de regresie în populația din care s-a extras proba se poate găsi între liniile ce unesc limitele inferioare și cele superioare, cu o probabilitate de $1-\alpha$.

Limitele de confidență ale unei estimări individuale

Estimarea unei valori $\hat{y}$ pentru o valoare individuală x care nu se regăsește în probă poate fi afectată de o sursă de eroare adițională, anume de împrăștierea față de dreaptă. Astfel, intervalul de confidență pentru o valoare va fi mai larg decât cel al dreptei de regresie. Intervalul se află folosind relațiile de mai sus, doar că se modifică formula erorii standard a unui punct de pe dreaptă prin adăugarea unei unități la valoarea dintre parantezele pătrate.

$$s_{\hat{y}} = \sqrt{\overline{SP}_{int}\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum x^2-\frac{(\sum x)^2}{n}}\right]}$$

Limita inferioară (LI) și cea superioară (LS) a intervalului de confidență pentru o valoare $\hat{y}$ corespunzătoare unei valori x din probă se calculează pornind de la relația:

$$\mu_{\hat{y}} = \hat{y} \pm s_{\hat{y}} \cdot t_{(\alpha,n-2)}$$
$$LI = \hat{y} - s_{\hat{y}} \cdot t_{(\alpha,n-2)}$$
$$LS = \hat{y} + s_{\hat{y}} \cdot t_{(\alpha,n-2)}\,.$$

Concluzia va fi că intervalul *LI-LS* pentru o valoare x include media populației de valori $\hat{y}$ ($\mu_{\hat{y}}$) cu o probabilitate de $1-\alpha$.

Unirea limitelor inferioare între ele și a celor superioare între ele, obținute pentru fiecare valoare x, duce la reprezentarea grafică a zonei de confidență pentru estimarea lui y pornind de la o valoare unică x (fig. 9.8).

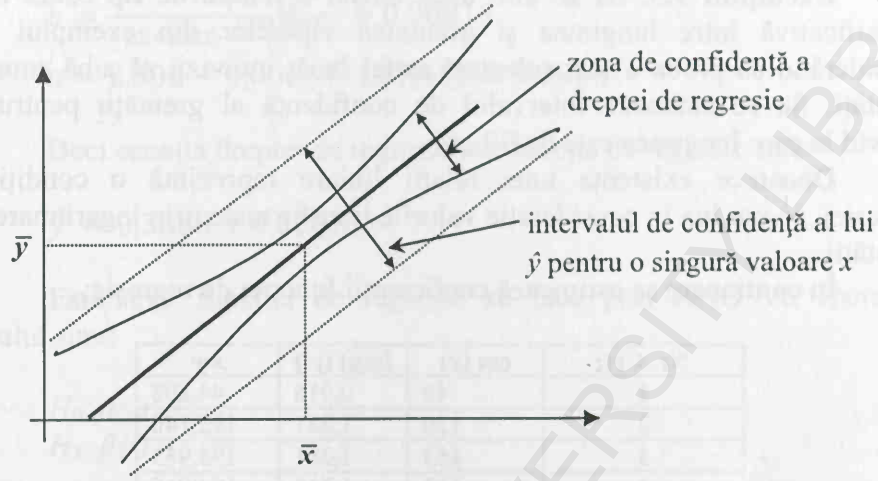


Figura 9.8. Zona de confidență a dreptei de regresie și pentru o valoare unică x

Abordarea relațiilor neliniare

Numeroase relații dintre variabile biologice nu sunt rectilinii. Un exemplu în acest sens îl constituie creșterea populațiilor și relația dintre mortalitate și vârstă.

În astfel de cazuri, relațiile curbilinii pot fi „îndreptate" prin transformarea datelor (secțiunea 4.5 – Transformarea datelor).

La început trebuie observat care din cele două variabile trebuie transformată. Aceasta se poate face prin încercări repetate, adică prin reprezentarea grafică a corelației cu câte o variabilă transformată sau cu amândouă.

O altă modalitate constă în calcularea r^2 pentru fiecare transformare alternativă ($r^2_{x,y}$; $r^2_{x',y}$; $r^2_{x,y'}$; $r^2_{x',y'}$). Transformarea care va duce la obținerea valorii celei mai mari a r^2 va fi folosită în analiza regresiei.

Transformarea variabilelor determină și modificarea funcției:

dacă $y \rightarrow y'$, atunci $a \rightarrow a'$ și $y' = a' + bx$;
dacă $y \rightarrow y'$ și $x \rightarrow x'$, atunci $y' = a' + bx'$.

Dacă se apelează la o transformare a datelor, se impune în final o transformare inversă a rezultatelor (coeficienți de regresie, limitele intervalelor de confidență).

Exemplul 9.3. Să se afle dacă există o relație de tip cauză-efect semnificativă între lungimea și greutatea viperelor din exemplul 9.1, considerând că proba a fost selectată astfel încât indivizii să aibă anumite greutăți. Să se estimeze intervalul de confidență al greutății pentru un individ la care lungimea este de 30 *cm*.

Deoarece existența unei relații liniare reprezintă o condiție a regresiei, se vor lua în considerație valorile transformate prin logaritmare ale greutății.

În continuare se estimează coeficienții funcției de regresie:

Nr. Crt.	cm (x)	ln(g) (y')	$x \cdot y$
1	49	0,916	44,898
2	150	1,281	192,140
3	153	1,281	195,983
4	146	1,335	194,910
5	164	1,386	227,352
6	66	1,504	99,269
7	160	1,548	247,610
8	215	2,230	479,453
9	228	2,351	536,114
10	222	2,389	530,305
11	260	2,639	686,155
12	325	3,045	989,470
13	340	3,091	1050,954
14	430	3,401	1462,515
15	390	3,512	1369,503
16	373	3,648	1360,725
17	425	3,689	1567,774
18	450	3,714	1671,107
19	389	3,835	1491,870
20	422	3,908	1649,182
21	418	3,976	1661,941
22	435	4,043	1758,727
23	459	4,279	1964,263
24	470	4,340	2039,754
Suma	7139	67,341	23471,976
Media	297,458	2,806	
$(\text{Suma})^2$	50965321	4534,757	
Suma de pătrate	2546985	218,368	

$$b = \frac{23471{,}976 - \frac{7139 \cdot 67{,}341}{24}}{2546985 - \frac{50965321}{24}} = 0{,}00813$$

$$a = 2{,}806 - 0{,}00813 \cdot 297{,}458 = 0{,}3886 .$$

Deci ecuația dreptei de regresie sau funcția de regresie este:

$$\hat{y}' = 0{,}3886 + 0{,}00813x .$$

Estimarea funcției de regresie se face prin ANOVA. Ipotezele testului sunt:

H_0: β=0
H_1: $\beta \neq 0$.

Se calculează sumele de pătrate:

$$SP_t = 218{,}368 - \frac{4534{,}757}{24} = 29{,}42$$

$$SP_{ext} = 0{,}00813\left(23471{,}976 - \frac{7139 \cdot 67{,}341}{24}\right) = 27{,}96$$

$$SP_{int} = 29{,}42 - 27{,}96 = 1{,}46 .$$

Se calculează numărul gradelor de libertate:

$$gl_t = 24 - 1 = 23$$
$$gl_{ext} = 2 - 1 = 1$$
$$gl_{int} = 24 - 2 = 22 .$$

Se calculează sumele de pătrate medii:

$$\overline{SP}_{ext} = \frac{27{,}96}{1} = 27{,}96 \qquad \overline{SP}_{int} = \frac{1{,}46}{22} = 0{,}0663 .$$

Se calculează statistica testului:

$$F = \frac{27{,}96}{0{,}0663} = 422{,}05 .$$

Se completează tabelul Anova:

Sursa de variabilitate	SP	gl	$\overline{SP}$	F
Externă	27,96	1	27,96	422,05
Internă	1,46	22	0,0663	
Totală	29,42	23		

Se află valoarea critică (anexa 2 sau 3):

$$F_{(0,05,1,22)} = 4,301 .$$

Statistica testului este mai mare decât valoarea critică, deci se respinge ipoteza nulă și se acceptă ipoteza alternativă pentru o probabilitate de 0,95. Probabilitatea calculată a ipotezei nule pentru statistica testului (anexa 3) este $7,6 \cdot 10^{-16}$.

Concluzia testului este că există o relație de tip cauză-efect semnificativă între cele două variabile, definită de funcția de regresie estimată.

Cu ajutorul funcției de regresie se pot calcula valorile estimate $\hat{y}'$ pentru valorile x. Prin punctele de coordonate $x, \hat{y}'$ se poate trasa dreapta de regresie. Valoarea $\hat{y}'$ pentru prima valoare $x = 49$ se calculează astfel:

$$\hat{y}' = 0,3886 + 0,00813 \cdot 49 = 0,7868 .$$

La fel se procedează și pentru celelalte valori ale variabilei x.

Pentru a construi zona de confidență a dreptei de regresie trebuie reprezentate grafic limitele intervalului de confidență ale mediei populației de valori y corespunzătoare fiecărei valori x. Pentru aceasta este nevoie să se afle valoarea critică $t_{(0,05,22)}$ (anexa 2 sau 3):

$$t_{(0,05,24-2)} = 2,074 .$$

Limitele intervalului de confidență pentru prima valoare $x = 49$ se calculează astfel:

$$LI' = 0{,}7868 - 2{,}074 \cdot \sqrt{0{,}0663 \cdot \left[\frac{1}{24} + \frac{(49-297{,}458)^2}{2546985-\frac{(7139)^2}{24}}\right]} = 0{,}5557$$

$$LS' = 0{,}7868 + 2{,}074 \cdot \sqrt{0{,}0663 \cdot \left[\frac{1}{24} + \frac{(49-297{,}458)^2}{2546985-\frac{(7139)^2}{24}}\right]} = 1{,}0179 .$$

La fel se procedează și pentru celelalte valori ale variabilei x.

cm (x)	*ln*(g) (y')	$\hat{y}'$	$s_{\hat{y}'}$	LI'	LS'
49	0,9163	0,7868	0,1114	0,5557	1,0179
150	1,2809	1,6076	0,0785	1,4448	1,7704
153	1,2809	1,6319	0,0776	1,4709	1,7929
146	1,3350	1,5751	0,0797	1,4098	1,7403
164	1,3863	1,7213	0,0745	1,5669	1,8758
66	1,5041	0,9249	0,1056	0,7060	1,1439
160	1,5476	1,6888	0,0756	1,5320	1,8456
215	2,2300	2,1358	0,0618	2,0075	2,2640
228	2,3514	2,2414	0,0593	2,1185	2,3644
222	2,3888	2,1927	0,0604	2,0673	2,3180
260	2,6391	2,5015	0,0546	2,3882	2,6147
325	3,0445	3,0297	0,0537	2,9184	3,1410
340	3,0910	3,1516	0,0552	3,0372	3,2660
430	3,4012	3,8829	0,0742	3,7290	4,0369
390	3,5115	3,5579	0,0640	3,4251	3,6907
373	3,6481	3,4197	0,0604	3,2944	3,5451
425	3,6889	3,8423	0,0728	3,6912	3,9934
450	3,7136	4,0455	0,0800	3,8795	4,2114
389	3,8351	3,5498	0,0638	3,4174	3,6821
422	3,9080	3,8179	0,0720	3,6686	3,9673
418	3,9759	3,7854	0,0710	3,6383	3,9326
435	4,0431	3,9236	0,0756	3,7667	4,0804
459	4,2794	4,1186	0,0827	3,9470	4,2902
470	4,3399	4,2080	0,0861	4,0294	4,3866

Pentru a afla intervalul de confidență al mediei populației de valori y pentru $x = 300$ se calculează relațiile:

$$\hat{y}' = 0{,}3886 + 0{,}00813 \cdot 300 = 2{,}8265$$

$$LI' = 2{,}8265 - 2{,}074 \cdot \sqrt{0{,}0663 \cdot \left[1 + \frac{1}{24} + \frac{(300-297{,}458)^2}{2546985 - \frac{(7139)^2}{24}}\right]} =$$

$$= 2{,}2295$$

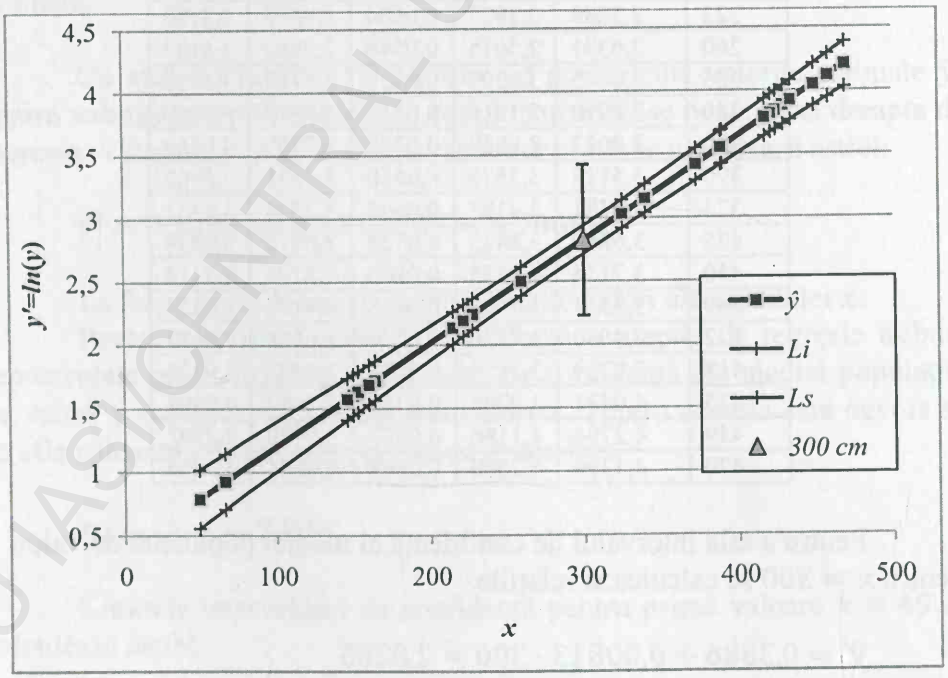
$$LS' = 2{,}8265 + 2{,}074 \cdot \sqrt{0{,}0663 \cdot \left[1 + \frac{1}{24} + \frac{(300-297{,}458)^2}{2546985 - \frac{(7139)^2}{24}}\right]} =$$

$$= 3{,}4235 .$$

Valorile obținute sunt modificate de transformarea inițială a greutăților. Pentru a obține valorile în grame este necesară transformarea inversă (anexa 3) a valorilor obținute:

$$\hat{y} = 16{,}8864 \qquad LI = 9{,}2956 \qquad LS = 30{,}6760 .$$

Concluzia este că o viperă de 300 *cm* poate avea o greutate între 9,3 *g* și 30,7 *g* cu o probabilitate de 0,95 sau în 95% din cazuri.



10. ANALIZA FRECVENȚELOR ȘI A DATELOR NOMINALE

Numeroase cercetări ecologice presupun numărarea și clasificarea lucrurilor folosind diferite scale nominale, cum ar fi speciile, culorile, habitatele etc. Din această cauză tehnicile statistice care analizează frecvențele sunt deosebit de utile. Metoda clasică de analiză a frecvențelor este testul χ^2. Statistica testului este comparată cu distribuția χ^2. Acesta este o distribuție a varianței probei. Distribuția χ^2 este asimetrică față de varianța populațională (σ^2). Partea stângă a distribuției ajunge la 0, în timp ce cea dreaptă poate atinge, teoretic, infinitul. Cu cât numărul gradelor de libertate crește, cu atât distribuția devine mai simetrică, iar în cazul probelor cu mai mult de 100 de unități de probă ($n > 100$) distribuția tinde să devină normală.

Împărțirea frecvențelor la numărul total de observații duce la transformarea distribuției frecvențelor într-o distribuție a probabilităților. Standardizarea axei orizontale prin înmulțirea varianței cu numărul gradelor de libertate și împărțirea produsului la varianța populației duce la obținerea distribuției χ^2.

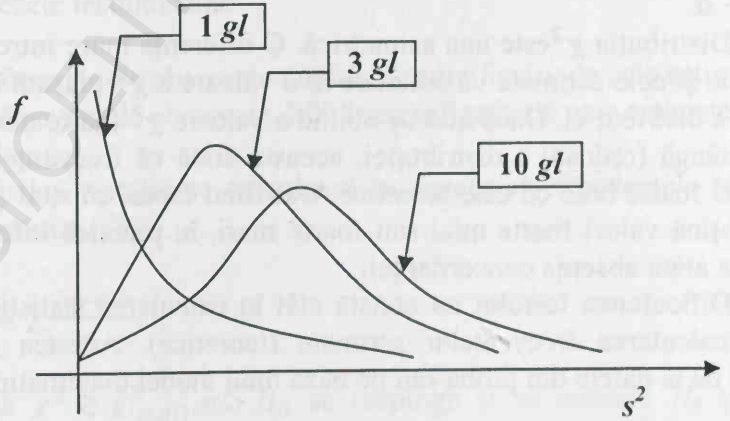


Figura 10.1. Distribuția varianței probei pentru 1, 3 și 10 grade de libertate

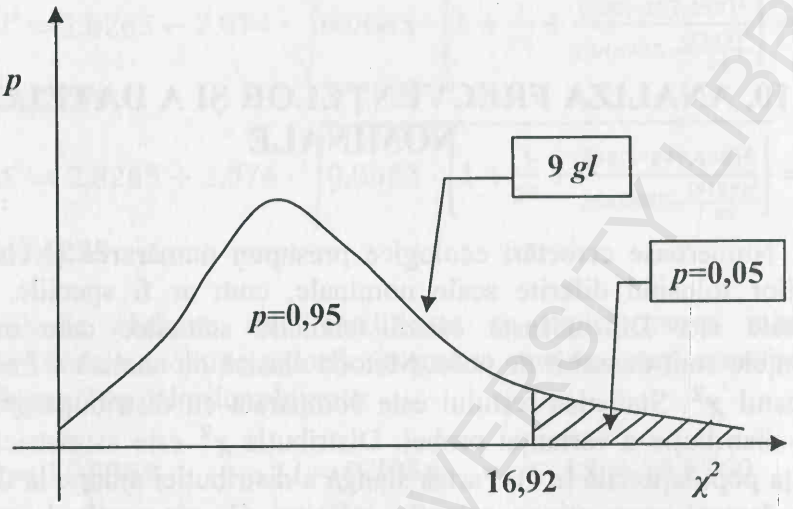


Figura 10.2. Distribuția χ^2 cu zona de respingere a ipotezei nule pentru o probabilitate de 0,05

Testele χ^2 pot fi pentru omogenitate, asociere, independență și de concordanță. Principiul acestor teste este același: frecvențele observate sunt comparate cu cele calculate teoretic sau estimate. Dacă între frecvențele observate și cele estimate există o diferență semnificativă, atunci statistica χ^2 va fi mai mare decât valoarea critică pentru gradele de libertate respective, caz în care H_0 se respinge, iar H_1 se acceptă în consecință cu $p = 1 - \alpha$.

Distribuția χ^2este una asimetrică. O diferență mare între frecvențele observate și cele estimate va conduce la o valoare a χ^2 plasată în coada din dreapta a distribuției. Dacă însă se obține o valoare χ^2 foarte mică plasată în coada stângă (redusă) a distribuției, aceasta arată că frecvențele observate concordă foarte bine cu cele teoretice. Dat fiind faptul că sunt șanse reduse să se obțină valori foarte mici sau foarte mari, în practică interesul constă doar în a arăta absența concordanței.

Dificultatea testului nu constă atât în calcularea statisticii acestuia, cât în calcularea frecvențelor estimate (teoretice). Acestea se pot afla pornind de la datele din probă sau pe baza unui model matematic.

10.1. TESTUL χ^2 DE CONCORDANȚĂ

Acesta verifică dacă distribuția frecvențelor observate în probe concordă sau nu într-o oarecare măsură cu o distribuție teoretică, cum ar fi Poisson, binomială sau normală, sau cu orice alt tip specificat de distribuție.

Testul de concordanță verifică cât de bine se potrivește un set de frecvențe observate aparținând la două sau mai multe categorii distincte cu o anumită distribuție așteptată. Testul este neparametric și necesită doar observații nominale sau sub formă de frecvențe.

Testul se folosește dacă datele îndeplinesc următoarele condiții:

1. Variabila trebuie să fie nominală; categoriile scalei nominale ale căror frecvențe sunt urmărite trebuie să nu se suprapună.
2. Observațiile sunt independente.
3. Frecvențele estimate trebuie să fie mai mari sau egale cu 5 sau, dacă există mai multe categorii, 80% dintre acestea trebuie să aibă frecvențele estimate mai mari sau egale cu 5.

Dacă a treia condiție nu este îndeplinită, categoriile cu frecvențele estimate mai mici de 5 se reunesc într-o singură categorie până când frecvența estimată a acesteia egalează sau depășește valoarea 5. Aceasta determină o reducere corespunzătoare a gradelor de libertate.

Ipotezele testului sunt:

H_0: frecvențele observate nu diferă semnificativ de cele estimate;
H_1: frecvențele observate diferă semnificativ de cele estimate.

Statistica testului se calculează în funcție de frecvențele observate (o) și cele estimate (e):

$$\chi^2 = \sum \frac{(o-e)^2}{e}.$$

Dacă $\chi^2 \geq \chi^2_{(\alpha,gl)} => H_0$ se respinge și se acceptă H_1 pentru o probabilitate de $1-\alpha$ $(100 \times (1-\alpha)\%)$.

$gl = k - m - 1$
k – numărul categoriilor (claselor de frecvență)
m – numărul parametrilor populaționali estimați.

Dacă frecvențele teoretice se obțin prin folosirea modelului Poisson, atunci $m = 1$ deoarece se estimează media populației necesară calculului probabilității în această distribuție (secțiunea 4.2), iar dacă se folosește modelul binomial sau binomial negativ, atunci $m = 2$ deoarece se estimează p și k (secțiunile 4.1 și, respectiv, 4.3).

Când există doar două categorii de distribuție, atunci vom avea 1 grad de libertate ($gl = 1$), iar statisticii testului i se aplică corecția Yates pentru continuitate. Aceasta înlătură posibilitatea obținerii unor valori prea mari ale statisticii testului. Corecția constă în scăderea valorii 0,5 din valoarea absolută a diferenței dintre frecvența observată și cea teoretică a fiecărei componente a formulei lui χ^2. Astfel, formula devine:

$$\chi^2 = \sum \frac{(|o-e|-0{,}5)^2}{e}.$$

Exemplul 10.1. Într-o probă extrasă aleator de pe fundul unui bazin acvatic au fost identificate 16 larve ale unei specii de chironomid. Larvele au fost crescute până la stadiul de adult, după care s-a determinat sexul indivizilor, rezultând 12 masculi și 4 femele. Există o diferență semnificativă între raportul dintre sexe (*sex ratio*) observat și cel de 1:1?

Sexul este o variabilă nominală ale cărei categorii se exclud reciproc. Observațiile sunt independente. Dacă *sex ratio* ar fi egal cu 1, atunci frecvențele estimate ale celor două sexe ar trebui să fie egale, adică din 16 indivizi, 8 ar trebui să fie masculi, iar 8 să fie femele. Putem concluziona că datele îndeplinesc condițiile de aplicare ale testului χ^2 de concordanță.

Ipotezele în cazul exemplului sunt următoarele:

H_0: *sex ratio* este egal cu 1, adică $masculi : femele = 1:1$;
H_1: *sex ratio* este diferit de 1.

	Frecvențe observate (o)	Frecvențe estimate (e)
Masculi	12	8
Femele	4	8
Total	16	16

Cum sunt doar două categorii de frecvențe (două valori ale variabilei) înseamnă că numărul gradelor de libertate va fi 1. Ca urmare, trebuie aplicată formula cu corecția pentru continuitate.

$$\chi^2 = \frac{(|12-8|-0,5)^2}{8} + \frac{(|4-8|-0,5)^2}{8} = 3,0625$$

Valoarea critică $\chi^2_{(0,05,1)} = 3,84$ (anexa 2 sau 3). Deoarece statistica testului este mai mică decât valoarea critică, nu se poate respinge ipoteza nulă pentru $p = 0,05$. Probabilitatea ca ipoteza nulă să fie adevărată (anexa 3) este 0,08, adică mai mare decât pragul de confidență de 0,05.

Concluzia testului este că nu există o deosebire semnificativă între *sex ratio* observat și cel teoretic de 1:1.

Exemplul 10.2. Pornind de la datele și rezultatele din exemplul 4.6 să se verifice semnificația concordanței dintre frecvențele estimate și cele teoretice.

Era vorba de densitatea indivizilor unei specii de șarpe din 100 de suprafețe de probă. Indicele de dispersie indică o dispersie aleatoare. S-au calculat frecvențele observate pe baza probabilităților estimate conform distribuției Poisson.

Dacă primele condiții de aplicare a testului χ^2 de concordanță sunt îndeplinite, se observă că ultima nu este îndeplinită – frecvențele estimate pentru valorile 6, 7 și 8 sunt mai mici decât valoarea 5. Pentru a putea continua testul se impune însumarea frecvențelor, atât observate, cât și estimate, corespunzătoare acestor valori. Ca urmare, frecvențele estimate și cele teoretice se prezintă astfel:

x	$f(o)$	$f'(e)$
0	7	5,727
1	16	16,379
2	25	23,422
3	18	22,329
4	16	15,965
5	10	9,132
6	5	4,353
7	2	1,778
8	1	0,636

x	o	e
0	7	5,727
1	16	16,379
2	25	23,422
3	18	22,329
4	16	15,965
5	10	9,132
6+7+8	8	6,767

Statistica testului se poate afla calculând valoarea raportului pentru fiecare pereche de frecvențe, după care se însumează.

$$\chi^2 = \frac{(7-5,727)^2}{5,727} + \frac{(16-16,379)^2}{16,379} + \frac{(25-23,422)^2}{23,422} + \frac{(18-23,329)^2}{23,329} + \\ + \frac{(16-15,965)^2}{15,965} + \frac{(10-9,132)^2}{9,132} + \frac{(8-6,767)^2}{6,767} = 1,544$$

Numărul gradelor de libertate se reduce datorită însumării ultimelor trei clase de frecvență. Astfel, din numărul de clase redus se scade 1 (a fost estimată media populațională prin media probei pentru calcularea funcției distribuției Poisson) și încă 1:

$$gl = 7 - 1 - 1 = 5 .$$

Valoarea critică în cazul acestui exemplu (anexa 2 sau 3) este $\chi^2_{(0,05,5)} = 11,07$. Deoarece statistica testului este mai mică decât valoarea critică, se acceptă ipoteza nulă pentru $p = 0,95$. Probabilitatea ca ipoteza nulă să fie adevărată (anexa 3) este 0,91, adică foarte mare.

Conform ipotezei nule nu există o diferență semnificativă între frecvențele observate și cele estimate conform distribuției Poisson.

Cum distribuția Poisson descrie probabilitatea de apariție a unor fenomene aleatoare, iar distribuția frecvențelor observate nu diferă semnificativ de această distribuție, se poate concluziona că distribuția indivizilor din populația cercetată este aleatoare, cu o probabilitate de 0,95.

În exemplul 4.6 spuneam că există o asemănare evidentă între frecvențele observate și cele estimate. Prin realizarea testului χ^2 de concordanță s-a completat rezolvarea problemei, în sensul că se poate specifica faptul că asemănarea sesizată anterior este semnificativă pentru un anumit nivel de încredere.

10.2. TESTUL χ^2 DE ASOCIERE

În analiza concordanței dintre frecvențele observate și cele calculate teoretic, era urmărită o singură variabilă a cărei valori apreciate pe o scală nominală reprezentau categorii de frecvență. Există însă situații în care se folosesc două variabile. De exemplu, un individ poate fi clasificat în funcție de sex și clasă de vârstă, specie și habitat etc. În astfel de cazuri, frecvențele se distribuie pe două sau mai multe rânduri, rezultând așa-numitul tabel sau matrice de contingență. Acestea permit investigarea asocierii dintre variabile. Unul dintre cele mai frecvent utilizate teste care verifică semnificația asocierii dintre variabilele apreciate pe o scală nominală este **testul χ^2 de asociere**. Acesta reprezintă echivalentul corelației pentru variabilele nominale.

Condiții de aplicare:

1. Datele trebuie să fie sub formă de frecvențe.
2. Observațiile sunt independente (o unitate de probă poate ocupa o singură poziție în matricea de contingență).
3. 80% dintre frecvențele estimate trebuie să fie mai mari sau egale cu 5 și nici o frecvență estimată să nu fie 0; deci, pentru o matrice de contingență de 2×2, toate celulele trebuie să aibă o valoare calculată teoretic mai mare sau egală cu 5.

Ipotezele testului sunt:

H_0: nu există asociere între variabile
H_1: există asociere între variabile.

Datele se aranjează într-un tabel cu două intrări, numit și matrice de contingență:

Variabila		**y (Coloană)**		**Total pe rând (TR)**
	Valori	**y_1 (C_1)**	**y_2 (C_2)**	
x (Rând)	**x_1 (R_1)**	a	b	$TR_1 = a + b$
	x_2 (R_2)	c	d	$TR_2 = c + d$
Total pe coloană (TC)		$TC_1 = a + c$	$TC_2 = b + d$	Total general (TG) $a + b + c + d$

Valorile teoretice se calculează pentru fiecare celulă în parte. De exemplu, valoarea estimată pentru prima celulă este:

$$e_{C_1R_1} = \frac{TR_1 \cdot TC_1}{TG}.$$

La fel se procedează și pentru celelalte celule ale matricei de contingență:

o	e
a	$\frac{(a+b)(a+c)}{a+b+c+d}$
b	$\frac{(a+b)(b+d)}{a+b+c+d}$
c	$\frac{(c+d)(a+c)}{a+b+c+d}$
d	$\frac{(c+d)(b+d)}{a+b+c+d}$

Statistica testului se calculează conform formulei:

$$\chi^2 = \sum \frac{(o-e)^2}{e}.$$

Dacă matricea de contingență are două linii și două coloane, atunci se utilizează corecția Yates pentru continuitate, la fel ca în cazul testului χ^2 de concordanță.

$$\chi^2 = \sum \frac{(|o-e|-0,5)^2}{e}$$

Dacă $\chi^2 \geq \chi^2_{(\alpha,(c-1)(r-1))} \Rightarrow H_0$ se respinge și se acceptă H_1 pentru o probabilitate de $1-\alpha$.

Când există mai multe categorii nominale ale celor două variabile analizate, atunci matricea de contingență poate avea mai mult de două rânduri și două coloane. În acest caz, metoda de calcul a frecvențelor estimate și a statisticii testului este aceeași cu cea folosită pentru o matrice de contingență cu două rânduri și două coloane ($r \times c = 2 \times 2$), cu excepția faptului că în formula de calcul a statisticii testului nu se mai aplică corecția Yates pentru continuitate.

Exemplul 10.3. Se crede ca femelele unei specii de șarpe acvatic dintr-un lac migrează toamna în apropierea bălților adiacente pentru depunerea pontei. Dacă este așa, atunci femelele ar trebui sa migreze mai intens decât masculii. Dacă din 27 de femele 25 au migrat și 2 nu au migrat, iar din 34 de masculi 4 au migrat, iar 30 nu, exista o asociere semnificativă între migrație și sex?

Datele sunt sub formă de frecvențe, iar șerpii pot aparține doar la una din cele două categorii definite de valorile variabilei nominale – sex. Deci primele două condiții de aplicare a testului sunt respectate de datele obținute în urma analizei probelor.

Se alcătuiește tabelul de contingență:

Variabila		**Migrația**		
	Valori	**Migratori**	**Nemigratori**	***TR***
Sex	**Femelă**	25	2	27
	Mascul	4	30	34
	TC	29	32	***TG***=61

Se calculează valorile estimate:

Sex	Migrația	o	e
Femelă	Migratori	25	(27×29)/61=12,8360
	Nemigratori	2	(27×32)/61=14,1639
Mascul	Migratori	4	(34×29)/61=16,1639
	Nemigratori	30	(34×32)/61=17,8360

Frecvențele teoretice (e) sunt toate mai mari ca 5, deci este îndeplinită și condiția a treia a testului χ^2 de asociere.

Ipotezele testului sunt:

H_0: nu există asociere semnificativă între migrație și sex
H_1: există o asociere semnificativă între migrație și sex.

Dat fiind că numărul gradelor de libertate este 1 ($(r-1)(c-1) = (2-1)(2-1)$) se aplică formula cu corecția pentru continuitate.

$$\chi^2 = \frac{(|25-12,8360|-0,5)^2}{12,8360} + \frac{(|2-14,1639|-0,5)^2}{14,1639} + \frac{(|4-16,1639|-0,5)^2}{16,1639} + \\ + \frac{(|30-17,8360|-0,5)^2}{17,8360} = 36,2484$$

Se află valoarea critică (anexa 2 sau 3): $\chi^2_{(0,05,1)} = 3,841$.

Statistica testului este mai mare decât valoarea critică și, în consecință, ipoteza nulă se respinge și se acceptă ipoteza alternativă pentru o probabilitate $p = 0,05$ (95%). Probabilitatea ca ipoteza nulă să fie adevărată este de $1,7 \cdot 10^{-9}$.

Se poate trage concluzia că există o asociere semnificativă între sex și migrație în populația de șerpi de apă analizată.

Exemplul 10.4. Într-un studiu s-a investigat preferința pentru habitat a larvelor aparținând la trei specii de insecte (A, B, C), prin prelevarea aleatoare a pupelor din trei ape curgătoare cu grade diferite de eutrofizare

(oligotrofică, mezotrofică și eutrofică). Există o asociere semnificativă între preferința pentru un anumit tip de habitat și apartenența specifică?

Datele sunt sub formă de frecvențe, iar o pupă poate aparține doar la o singură specie și poate proveni doar dintr-o singură apă curgătoare. Deci primele două condiții ale testului χ^2 de asociere sunt îndeplinite.

Se scrie matricea de contingență 3×3.

Variabila		**Habitat**			
	Valori	**Oligotrofic**	**Mezotrofic**	**Eutrofic**	***TR***
Specia	***A***	10	12	35	57
	B	7	26	11	44
	C	28	13	9	50
	TC	45	51	55	***TG***=151

Se calculează valorile estimate:

sp.	**Habitat**	***o***	***e***
A	**Oligotrofic**	10	(45×57)/151=16,9868
	Mezotrofic	12	(51×57)/151=19,2517
	Eutrofic	35	(55×57)/151=20,7614
B	**Oligotrofic**	7	(45×44)/151=13,1126
	Mezotrofic	26	(51×44)/151=14,8609
	Eutrofic	11	(55×44)/151=16,0265
C	**Oligotrofic**	28	(45×50)/151=14,9007
	Mezotrofic	13	(51×50)/151=16,8874
	Eutrofic	9	(55×50)/151=18,2119

Frecvențele teoretice (e) sunt toate mai mari ca 5, deci este îndeplinită și condiția a treia a testului χ^2 de asociere.

Ipotezele testului sunt:

H_0: nu există asociere semnificativă între apartenența specifică și tipul de habitat

H_1: există o asociere semnificativă între apartenența specifică și tipul de habitat.

Numărul gradelor de libertate este 4 ($(r-1)(c-1) = (3-1)(3-1)$). Astfel, se aplică formula statisticii testului fără corecția pentru continuitate.

$$\chi^2 = \frac{(10-16{,}9868)^2}{16{,}9868} + \frac{(12-19{,}2517)^2}{19{,}2517} + \frac{(35-20{,}7614)^2}{20{,}7614} + \frac{(7-13{,}1126)^2}{13{,}1126} + \\ + \frac{(26-14{,}8609)^2}{14{,}8609} + \frac{(11-16{,}0265)^2}{16{,}0265} + \frac{(28-14{,}9007)^2}{14{,}9007} + \frac{(13-16{,}8874)^2}{16{,}8874} + \\ + \frac{(9-18{,}2119)^2}{18{,}2119} = 45{,}2155$$

Se află valoarea critică (anexa 2 sau 3): $\chi^2_{(0{,}05,4)} = 9{,}488$.

Valoarea statisticii testului este mai mare decât valoarea critică, ipoteza nulă se respinge și se acceptă ipoteza alternativă pentru o probabilitate $p = 0{,}05$ (95%). Probabilitatea ca ipoteza nulă să fie adevărată este $8{,}3 \cdot 10^{-10}$, deci foarte mică.

Se poate trage concluzia că există o asociere semnificativă între apartenența specifică și gradul de eutrofizare al habitatelor.

Pentru a determina care dintre speciile analizate este asociată cu un anumit grad de eutrofizare, trebuie comparate valorile observate cu cele estimate (așteptate, dacă nu ar fi existat nici o preferință pentru habitat): pentru specia A, frecvența observată a fost mai mare decât cea estimată în apele eutrofe; pentru specia B, frecvența observată a depășit-o pe cea estimată în apele mezotrofe; pentru specia C, frecvența observată a fost mai mare decât cea estimată în apele oligotrofe.

Din acest exemplu se poate trage și o concluzie de ordin general: unde frecvențele observate sunt mai mari decât cele estimate înseamnă că există o asociere pozitivă între cele două variabile și invers, unde frecvențele estimate sunt mai mari decât cele observate înseamnă că există o asociere negativă între variabile.

10.3. TESTUL EXACT AL LUI FISHER

Este asemănător cu testul χ^2 pentru asociere și se folosește pentru testarea asocierii dintre două variabile nominale. Diferența dintre cele două teste constă în faptul că testul Fisher nu prevede a treia condiție de aplicare a testul χ^2. Primele două condiții (privind datele sub formă de frecvențe și independența probelor) trebuie îndeplinite. Acest test este foarte util pentru analiza probelor de dimensiuni mici. Un alt avantaj față de testul χ^2 se referă la faptul că testul Fisher se poate aplica și în variantă unilaterală.

Pentru calcularea testului, datele trebuie ordonate într-o matrice de contingență:

Valoarea variabilei		x		**Total**
		prezentă (C_1)	**absentă (C_2)**	**Rând (TR)**
y	**prezentă (R_1)**	a	b	$a+b$
	absentă (R_2)	c	d	$c+d$
Total	**coloană (TC)**	$a+c$	$b+d$	$TG = a+b+c+d$

Ipotezele testului sunt:

Bilateral	**Unilateral**
H_0: nu există asociere între x și y.	H_0: indivizii care prezintă x nu au tendința să prezinte y, iar cei care nu prezintă x au tendința să prezinte y.
H_1: există asociere între x și y.	H_1: indivizii care prezintă x au tendința să prezinte și y, iar cei care nu prezintă x au tendința să nu prezinte nici y.

Deosebirea între ipotezele bilaterale și cele unilaterale, în cazul testului Fisher, este similară cu cea dintre aceleași ipoteze în situația unei analize a corelației: ipotezele bilaterale ale testului Fisher sunt analoage

ipotezelor privind nonexistența sau existența unei asocieri între două variabile, fără a preciza tipul corelației; ipotezele unilaterale ale testului Fisher sunt similare ipotezelor care specifică tipul corelației (pozitivă sau negativă).

Se scriu apoi toate matricele posibile pentru aceleași totaluri marginale, adică se modifică frecvențele a, b, c, și d astfel încât TC_1, TC_2, TR_1, TR_2 și TG să rămână constante. Astfel, rezultă o serie de matrice în care la un capăt poziția a va avea o valoare minimă, iar la celălalt va avea o valoare maximă.

Se calculează disproporționalitatea pentru fiecare matrice obținută, după relația:

$$d = \left|\frac{a}{a+b} - \frac{c}{c+d}\right|.$$

Se calculează probabilitatea (p) pentru fiecare matrice a cărei disproporții (d) este mai mare sau egală cu disproporția matricei inițiale, conform relației

$$p = \frac{TC_1! \cdot TC_2! \cdot TR_1! \cdot TR_2!}{TG! \cdot a! \cdot b! \cdot c! \cdot d!}.$$

Suma probabilităților acestor matrice reprezintă probabilitatea ca H_0 să fie adevărată. Se respinge ipoteza nulă dacă suma probabilităților este mai mică decât pragul de semnificație (0,05).

Dacă testul se aplică în varianta bilaterală, atunci se însumează probabilitățile date de toate matricele din ambele capete ale seriei cu disproporționalitatea mai mare sau egală cu cea a matricei inițiale. Când se formulează ipoteze unilaterale, atunci se însumează doar probabilitățile matricelor dintr-o singură parte a seriei, ale căror disproporționalitate este mai mare sau egală cu cea a matricei inițiale.

Exemplul 10.5. Într-un experiment s-a urmărit dacă juvenilii unei specii de șarpe tind să prezinte o reacție de fugă la un stimul amenințător de deasupra sau din lateral. Ca urmare, 7 șerpi au fost stimulați de deasupra capului, iar alți 7, din lateral și s-au înregistrat tentativele de scăpare.

În această situație este vorba de două probe independente, deoarece un individ a fost supus unui singur tratament (a fost stimulat fie de deasupra, fie din lateral), și de categorii care se exclud reciproc (au prezentat reacția sau nu au prezentat-o). În aceste condiții, pentru a realiza testarea, datele trebuie aranjate într-un tabel de contingență. Acesta poate fi alcătuit astfel încât să reflecte dacă se urmărește o ipoteză bilaterală sau unilaterală:

Tabel de contingență pentru varianta bilaterală

	Reacție prezentă	Reacție absentă
Stimul de deasupra	6	1
Stimul din lateral	1	6

Tabel de contingență pentru varianta unilaterală

		Reacție de scăpare	
		da	Nu
Stimul de deasupra capului	da	6	1
	nu	1	6

Dacă se calculează frecvențele estimate conform testului χ^2 de asociere, se observă că toate acestea sunt mai mici decât 5 (frecvențele estimate sunt egale cu 3,5), deci condiția a treia a testului nu este îndeplinită. Ca urmare, testul χ^2 de asociere nu poate fi aplicat pentru acest exemplu.

Ipotezele testului:

Bilateral	**Unilateral**
H_0: nu există asociere între stimulare și apariția reacției.	H_0: indivizii care au fost stimulați de deasupra nu tind să prezinte reacția de scăpare.
H_1: există asociere între stimulare și apariția reacției.	H_1: indivizii care au fost stimulați de deasupra tind să manifeste o reacție de scăpare.

Indiferent de varianta în care se aplică testul, trebuie scrisă seria de matrice. Cel mai simplu este ca la valoarea din poziția a să se adune câte o unitate până se ajunge la capătul din dreapta al seriei, iar apoi din aceeași

valoare să se scadă câte o unitate până se obține capătul din stânga al seriei. În funcție de valoarea a și totalurile marginale se completează celelalte poziții ale matricelor.

#1	
0	*7*
7	*0*
$d = 1$	

#2	
1	*6*
6	*1*
$d = 0,71$	

#3	
2	5
5	2
$d = 0,42$	

#4	
3	4
4	3
$d = 0,14$	

#5	
4	3
3	4
$d = 0,14$	

#6	
5	2
2	5
$d = 0,42$	

#7	
6	**1**
1	**6**
$d = 0,71$	

#8	
7	*0*
0	*7*
$d = 1$	

Matricele care au d mai mare sau egal cu d matricei inițiale (#7) sunt matricele #1, #2 și #8. Deci pentru acestea se vor calcula probabilitățile:

$$p(\#1) = \frac{7!\cdot 7!\cdot 7!\cdot 7!}{14!\cdot 0!\cdot 7!\cdot 7!\cdot 0!} = 0,00029$$
$$p(\#2) = \frac{7!\cdot 7!\cdot 7!\cdot 7!}{14!\cdot 1!\cdot 6!\cdot 6!\cdot 1!} = 0,01427$$
$$p(\#8) = \frac{7!\cdot 7!\cdot 7!\cdot 7!}{14!\cdot 7!\cdot 0!\cdot 0!\cdot 7!} = 0,00029\,.$$

Pentru a calcula probabilitatea ipotezei nule este nevoie și de probabilitatea dată de matricea inițială:

$$p(\#7) = \frac{7!\cdot 7!\cdot 7!\cdot 7!}{14!\cdot 6!\cdot 1!\cdot 1!\cdot 7!} = 0,00029\,.$$

Dacă testul se aplică în variantă **bilaterală**, atunci probabilitatea ipotezei nule va fi dată de suma probabilităților celor patru matrice:

$$p(H_0) = p(\#1) + p(\#2) + p(\#7) + p(\#8)$$
$$p(H_0) = 0,00029 + 0,01427 + 0,01427 + 0,00029 = 0,0291\,.$$

Dacă testul se aplică în variantă **unilaterală**, atunci probabilitatea ipotezei nule va fi suma probabilităților calculate pe baza matricelor #7 și #8:

$$p(H_0) = p(\#7) + p(\#8)$$
$$p(H_0) = 0{,}01427 + 0{,}00029 = 0{,}0145\,.$$

Indiferent de varianta aplicată, condiția testului este următoarea:

dacă $p(H_0) \leq 0{,}05 \Rightarrow H_0$ se respinge și se acceptă H_1.

În cazul exemplului, ambele rezultate arată că ipotezele nule corespunzătoare celor două variante au o probabilitate foarte mică (mai mică de 0,05). Deci se poate accepta fie că există o asociere semnificativă între stimulare și reacția de scăpare, fie că există o asociere semnificativă între stimularea de deasupra capului și apariția reacției de scăpare. Evident, al doilea rezultat (varianta unilaterală) este în cazul experimentului prezentat mai valoros, deoarece oferă un plus de informație.

10.4. TESTUL MCNEMAR

În anumite situații nu se dorește ca observațiile să fie independente (de exemplu, dacă se investighează acțiunea unui tratament asupra unor animale). Astfel, pentru a elimina diferențele individuale, aceiași indivizi vor fi investigați de două ori – înainte și după tratament.

În astfel de cazuri se folosește testul McNemar pentru testarea semnificației schimbărilor răspunsurilor organismelor sub influența unui tratament. Acest test este echivalentul testului t pentru perechi de valori în cazul variabilelor nominale.

Condiții de aplicare:

1. variabila trebuie să fie nominală;
2. observațiile trebuie să fie neindependente (fiecare individ trebuie să fie investigat de două ori).

Se alcătuiește matricea de contingență:

Tratament		Tratament 1	
	Răspuns	Răspuns 1	Răspuns 2
Tratament 2	Răspuns 1	a	b
	Răspuns 2	c	d

Ipotezele testului sunt următoarele:

H_0: răspunsul nu se modifică semnificativ sub acțiunea tratamentului
H_1: răspunsul se modifică semnificativ sub acțiunea tratamentului

Indivizii care și-au modificat comportamentul ocupă pozițiile b și c (indivizii din b au dat răspunsul 2 la tratamentul 1 și răspunsul 1 la tratamentul 2; indivizii din c au dat răspunsul 1 la tratamentul 1 și răspunsul 2 la tratamentul 2). Indivizii din pozițiile a și d nu și-au modificat răspunsul (indiferent de tratament ei au dat fie răspunsul 1, fie 2).

Statistica testului consideră doar indivizii care și-au modificat răspunsul în funcție de tratament:

$$\chi^2 = \frac{(|c-b|-1)^2}{c+b}.$$

Statistica testului de compară cu o valoare critică aleasă (anexa 2 sau 3) în funcție de pragul de confidență și numărul gradelor de libertate.

Dacă $\chi^2 \geq \chi^2_{(\alpha,(c-1)(r-1))} \Rightarrow H_0$ se respinge și se acceptă H_1 pentru $p = 1 - \alpha$.

Dacă $\frac{c+b}{2} < 5$, se poate calcula probabilitatea H_0 cu ajutorul funcției distribuției binomiale. În acest caz parametrii distribuției se calculează astfel:

$$k = c + b \qquad p = 0{,}5 \qquad x = min(c, b)$$
$$p(x) = \frac{k!}{x!(k-x)!} \cdot p^x \cdot q^{(k-x)}.$$

Dacă $p(x) < 0,05 \Rightarrow H_0$ se respinge și se acceptă H_1 pentru $p = 1 - 0,05 = 0,95$.

Exemplul 10.6. La unele specii de amfibieni anuri, la care dimorfismul sexual nu este evident, din repertoriul masculilor fac parte printre altele și sunetele de eliberare. Acestea sunt emise de un mascul atunci când este abordat de un altul (care îl confundă cu o femelă) pentru a realiza împerecherea (amplexus sau îmbrățișare). Astfel, masculul abordat semnalizează sonor că și el este tot mascul. În acest sens, într-un studiu efectuat pe 31 de masculi de *Pelophylax* (*Rana*) *lessonae*, s-a aplicat un stimul tactil în zona axilară (zona în care masculul prinde femela în amplexus) și apoi un altul în zona spatelui. Rezultatele au fost următoarele: 25 au reacționat prin emiterea sunetului de eliberare doar în urma stimulării axilare, 2 au emis sunetul când au fost stimulați doar dorsal, 3 au emis sunetul în urma ambelor stimulări și 1 nu a reacționat la niciunul dintre stimuli. Există o modificare semnificativă a comportamentului în funcție de zona în care este aplicat stimulul?

În cazul acestui experiment, nu se poate aplica testul χ^2 de asociere și nici testul Fisher, deoarece aplicarea ambelor teste este condiționată de existența unor observații independente, iar în acest experiment același individ a fost supus ambelor tratamente: o dată a fost stimulat axilar, iar a doua oară, dorsal. Deci pentru testarea semnificației asocierii dintre comportament și zona stimulată trebuie aplicat testul McNemar.

Matricea de contingență în acest caz este următoarea:

Zona stimulată		Axilară	
	Sunet de eliberare	Prezent	Absent
Dorsală	Prezent	3	2
	Absent	25	1

Ipotezele testului sunt:

H_0: nu există o schimbare semnificativă a comportamentului în funcție de zona stimulată.

H_1: există o schimbare semnificativă a comportamentului în funcție de zona stimulată.

În matricea de contingență, masculii care și-au schimbat comportamentul în funcție de zona stimulată sunt cei din pozițiile b (2 au reacționat la stimulul dorsal și nu au reacționat la cel axilar) și c (25 au reacționat la stimulul axilar și nu au reacționat la cel dorsal). Cei din pozițiile a și d nu și-au schimbat comportamentul (3 au reacționat la ambele stimulări și, respectiv, 1 nu a reacționat la nici un stimul).

Pe baza acestor date se calculează statistica χ^2 a testului McNemar:

$$\chi^2 = \frac{(|25-2|-1)^2}{25+2} = 17{,}926\,.$$

Se află valoarea critică (anexa 2 sau 3) pentru $\alpha = 0{,}05$ și 1 grad de libertate (tabelul are 2 rânduri și 2 coloane, deci $(2-1)(2-1) = 1$).

$$\chi^2_{(0{,}05,1)} = 3{,}841\,.$$

Valoarea statisticii testului este mai mare decât valoarea critică, deci se respinge ipoteza nulă și se acceptă ipoteza alternativă. Probabilitatea ca ipoteza nulă să fie adevărată (anexa 3) este foarte mică ($2{,}3 \cdot 10^{-5}$).

Concluzia testului este că există o schimbare semnificativă a comportamentului ce constă în emiterea sunetului de eliberare în funcție de zona stimulată la masculii speciei investigate.

Să presupunem că rezultatul experimentului mai sus menționat se prezintă astfel:

Zona stimulată		Axilară	
	Sunet de eliberare	Prezent	Absent
Dorsală	Prezent	3	0
	Absent	9	1

În acest caz, media dintre b și c este 4,5 (o valoare mai mică decât 5) și, ca urmare, statistica testului nu mai este aproximată de distribuția χ^2. În această situație se folosește un test exact bazat pe distribuția binomială, cu $p = q = 0{,}5$ și $k = c + b$.

Parametrii necesari calculării probabilității sunt: $x = 0$; $p = q = 0{,}5$; $k = 9 + 0 = 9$.

$$p(0) = \frac{9!}{0!(9-0)!} \cdot 0{,}5^0 \cdot 0{,}5^{(9-0)} = 0{,}0019$$

Valoarea obținută este mai mică decât 0,05 sau probabilitatea ipotezei nule este 0,0019, deci putem accepta ipoteza alternativă conform căreia există o schimbare semnificativă a comportamentului în funcție de locul unde s-a aplicat stimulul.

BIBLIOGRAFIE

Andrei T., Stancu S. (1995): *Statistica – teorie şi aplicaţie*. Editura All.

Armitage P., Colton T. (2005): *Encyclopedia of Biostatistics*, 2nd edition. John Wiley and Sons, Ltd.

Bailey T.J.N. (1981): *Statistical Methods in Biology*, 2nd edition. Cambridge University Press.

Bart J., Fligner M.A., Notz W.I. (2004): *Sampling and statistical methods for behavioral ecologists*. Cambridre University Press.

Bennett P.D., Humphries A.D. (1977): *Introduction to Field Biology*. Edward Arnold (Publishers) Ltd.

Bishop O.N. (1971): *The Principles of Modern Biology – Statistics for Biology*, 2nd edition. Longman.

Cann A.J. (2002): *Maths from Scratch for Biologists*. John Wiley & Sons, Ltd.

Ceapoiu M. (1968): *Metode statistice aplicate în experienţele agricole şi biologice*. Editura Agro-Silvică, Bucureşti.

Cox W.G. (1996): *Laboratory Manual of General Ecology*, 7th edition. Wm. C. Brown Publishers.

Dragomirescu L. (1998): *Biostatistică pentru începători*. Editura Constelaţii, Bucureşti.

Dragomirescu L. (1999): *Lucrări practice de biostatistică*. Editura Ars Docendi, Bucureşti.

Dytham C. (2003): *Choosing and Using Statistics: A Biologist's Guide*, 2nd edition. Blackwell Publishing.

Everitt B.S. (2002): *The Cambridge Dictionary of Statistics*, 2nd edition. Cambridge University Press.

Forthofer R.N., Lee E.S., Hernandez M. (2007): *Biostatistics: A Guide to Design, Analysis, and Discovery*. Elsevier Inc.

Fowler J., Cohen L., Javris P. (2000): *Practical Statistics for Field Biology*, 2nd edition. John Wiley and Sons, Ltd.

Glantz S.A. (2005): *Primer of Biostatistics*, 6th edition. McGraw-Hill.

Glaser A.N. (2001): *High-Yeld™ Biostatistics*. Lippincott Williams & Wilkins.

Hampton E.R. (1994): *Introductory Biological Statistics*. Wm. C. Brown Publishers.

Härdle W., Mori Y., Vieu P. (2007): *Statistical Methods for Biostatistics and Related Fields*. Springer-Verlag Berlin Heidelberg.

Iosifescu M., Moineagu C., Trebici V., Ursianu E. (1985): *Mica enciclopedie de statistică*. Editura Ştiinţifică şi Enciclopedică, Bucureşti.

Le C.T. (2003): *Introductory Biostatistics*. John Wiley and Sons, Ltd.

Ludwig J.A., Reynolds J.F. (1988): *Statistical Ecology: A primer on Methods and Computing*. John Wiley and Sons, Ltd.

Manly B.F.J., McDonald L.L., Thomas D.L., McDonald T.L., Erikson W.P. (2004): *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*, 2nd edition. Kluwer Academic Publishers.

Michelson S., Schofield S. (2002): *The Biostatistics Cookbook: The most user-friendly guide for the bio/medical scientist*. Kluwer Academic Publishers.

Morisita M. (1962): I_δ-Index, a measure of dispersion of individuals. *Res. Popul. Ecol.*, 4: 1-7.

Neacşu P. (1987): *Lucrări practice de ecologie*. Bucureşti.

Neacşu P., Apostolache-Stoicescu Z. (1982): *Dicţionar de ecologie*. Editura Ştiinţifică şi Enciclopedică, Bucureşti.

Norman G.R., Streiner D.L. (1998): *Biostatistics: The Bare Essentials*. B.C. Decker Inc.

Pârvu C. (1999): *Ecologie generală*. Editura Tehnică, Bucureşti.

Petrie A., Sabin C. (2000): *Medical Statistics at a Glance*. Blackwell Science Ltd.

Postelnicu V., Coatu S. (1980): *Mica enciclopedie matematică* (traducere după Kleine Enzyklopadie der Mathematik ed. VI-a, 1971 şi Mathematics at a glance, 1975). Editura Tehnică, Bucureşti.

Simionescu V. (1983): *Lucrări practice de ecologie*. Editura Universităţii „Al.I. Cuza" Iaşi.

Slingsby D., Cook C. (1992): *Practical Ecology*. Macmillan Distribution Ltd.

Smith R.L. (1996): *Ecology and Field Biology*, 5th edition. Harper Collins College Publishers.

Snedecor W.G. (1968): *Metode statistice aplicate în cercetările de agricultură şi biologie* (traducere din limba engleză). Bucureşti.

Southwood T.R.E. (1966): *Ecological Methods with Particular Reference to the Study of Insect Populations*. London, Methuen and co. LTD.

Stan Gh. (1994): Metode statistice cu aplicaţii în cercetările entomologice (IV). *Bul. Inf. Soc. Lepid. Rom.*, 5 (1): 13-25.

Stan Gh., 1994, Metode statistice cu aplicaţii în cercetările entomologice (V). *Bul. Inf. Soc. Lepid. Rom.*, 5 (2): 113-126.

Stan Gh. (1994): Metode statistice cu aplicaţii în cercetările entomologice (VI). *Bul. Inf. Soc. Lepid. Rom.*, 5 (3-4): 257-280.

Stan Gh. (1995): Metode statistice cu aplicaţii în cercetările entomologice (VII). *Bul. Inf. Soc. Lepid. Rom.*, 6 (1-2): 67-96.

Stiling, P.D. (2001): *Ecology Theories and Applications*, 4th edition. Prentice Hall.

Varvara M. (2000): *Curs de Ecologie*, vol. 1. Editura Universităţii „Al.I. Cuza" Iaşi.

Varvara M., Zamfirescu Ş.R., Neacşu V. (2001): *Lucrări practice de ecologie – manual*. Editura Universităţii „Al.I. Cuza" Iaşi.

Zamfirescu O., Zamfirescu Ş.R. (2007): Aspects Regarding the Vegetation From the Floristic Reserve "The Secular Hayfields From Valea Lui David" Iaşi, România. *Journal of Ecology and Safety, International Scientific Publication*, 1:32-39.

Zamfirescu Ş.R. (2002): The Experimental Induction of the Release Calls of Some Anuran Species (Amphibia, Anura). In: Tomescu, N., Popa, V. (eds.), *In Memoriam "Prof. Dr. Doc. Vasile Ghe. Radu" Coresponding Member of Romanian Academy of Sciences*. Cluj University Press, Cluj-Napoca, pp. 169-172.

Zamfirescu Ş.R., Zamfirescu O., Popescu I.E., Ion C., Strugariu A. (2008): *Vipera de stepă (Vipera ursinii moldavica) şi habitatele sale din Moldova (Romania)*. Editura Universităţii „Al.I. Cuza" Iaşi.

ANEXA 1: CHEIE DIHOTOMICĂ PENTRU DETERMINAREA TIPULUI DE ANALIZĂ STATISTICĂ

1. Aprecierea datelor:
   a. datele sunt apreciate pe o scală ordinală, de interval, de raport → **2**
   b. datele sunt apreciate pe o scală nominală și/sau sub formă de frecvențe → **15**

2. Numărul variabilelor analizate:
   a. o singură variabilă (ex: lungimea, greutatea, număr de indivizi) → **3**
   b. două variabile (ex: lungimea și greutatea) → **13**

3. Numărul probelor analizate:
   a. o singură probă → **4**
   b. mai multe probe → **5**

4. Scopul analizei:
   a. descrierea tendinței centrale și a variabilității probei → **Statistica descriptivă.**
   b. compararea mediei cu o valoare control → **Testul Student (t) pentru o probă.**

5. Numărul probelor:
   a. 2 probe 6 (Teste pentru 2 probe) →**6**
   b. 3 sau mai multe probe 9 (ANOVA) →**9**

6. Independența observațiilor:
   a. observații independente (probele provin din populații diferite) → **7**
   b. observații neindependente (probele provin din aceeași populație sau sunt obținute prin efectuarea unor observații repetate asupra acelorași unități de probă) → **8**

7. Distribuția valorilor și scala de apreciere a variabilei:
   a. distribuție aproximativ normală, scală de interval sau de raport → **Testul Student (t) pentru observații independente.**
   b. distribuție diferită de cea normală, scală de interval, de raport sau ordinală → **Testul Mann-Whitney (U).**

8. Distribuția valorilor și scala de apreciere a variabilei:
   a. distribuție aproximativ normală, scală de interval sau de raport → **Testul Student (t) pentru observații perechi.**
   b. distribuție diferită de cea normală, scală de interval, de raport sau ordinală → **Testul Wilcoxon (W) pentru observații perechi.**

9. Numărul factorilor (tratamente) care influențează probele:
   a. 1 singur factor (tratament) → **10**
   b. 2 factori (tratamente) → **11**

10. Distribuția valorilor și scala de apreciere a variabilei:
    a. distribuție aproximativ normală, scală de interval sau de raport, varianța internă omogenă → **ANOVA unifactorială.**
    b. distribuție diferită de cea normală, scală de interval, de raport sau ordinală, varianță internă heterogenă → **ANOVA unifactorială neparametrică Kruskal-Wallis.**

11. Numărul de observații în celulă, interacțiunea dintre factori:
    a. o singură observație în celulă, nu există interacțiune → **12**
    b. mai multe observații în celulă, există interacțiune → **ANOVA bifactorială cu replicare.**

12. Distribuția valorilor și scala de apreciere a variabilei:
    a. distribuție aproximativ normală, scală de interval sau de raport, varianță internă omogenă → **ANOVA bifactorială fără replicare.**
    b. distribuție diferită de cea normală, scală de interval, de raport sau ordinală, varianță internă heterogenă → **ANOVA bifactorială neparametrică Friedman.**

13. Relația dintre variabile, prelevarea probei:
    a. asociere liniară, proba este prelevată aleator → **14** (Analiza Corelației)
    b. relație liniară cauză-efect, proba este prelevată în funcție de valorile variabilei independente → **Analiza Regresiei.**

14. Distribuția valorilor și scala de apreciere a variabilei:
    a. distribuție aproximativ normală, scală de interval sau de raport → **Corelația parametrică (Coeficientul de corelație Pearson).**
    b. distribuție diferită de cea normală, scală de interval, de raport sau ordinală → **Corelația neparametrică (Coeficientul de corelație Spearman).**

15. Scopul analizei:
    a. concordanța dintre distribuția frecvențelor observate și cea a frecvențelor estimate conform unei distribuții teoretice, cunoscute → **Testul Chi-Pătrat de concordanță.**
    b. asocierea dintre 2 variabile apreciate pe o scală nominală sau sub formă de frecvențe → **16**

16. Independența observațiilor:
    a. observații independente (fiecare unitate de probă este investigată o singură data) → **17**
    b. observații neindependente (sunt obținute prin efectuarea unor investigații repetate asupra aceloraşi unități de probă) → **Testul McNemar pentru semnificația schimbării.**

17. Magnitudinea frecvențelor estimate:
    a. frecvențele estimate mai mari ca 0 și cel mult 20% dintre ele sunt mai mici ca 5 → **Testul Chi-Pătrat de Asociere.**
    b. există frecvențe estimate egale cu 0 și mai mult de 20% din frecvențele estimate sunt mai mici ca 5 → **Testul exact al lui Fisher.**

ANEXA 2: TABELE CU VALORI CRITICE

Scorul z și probabilitățile corespunzătoare în distribuția normală standard

z	A doua zecimală									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Valorile critice ale distribuției *t* (Student)

Grade de libertate	α (bilateral)	
	0,05	0,1
	α (unilateral)	
	0,025	0,05
1	12,706	6,314
2	4,303	2,920
3	3,182	2,353
4	2,776	2,132
5	2,571	2,015
6	2,447	1,943
7	2,365	1,895
8	2,306	1,861
9	2,262	1,833
10	2,228	1,812
11	2,201	1,796
12	2,179	1,782
13	2,160	1,771
14	2,145	1,761
15	2,131	1,753
16	2,120	1,746
17	2,110	1,740
18	2,101	1,734
19	2,093	1,729
20	2,086	1,725
21	2,080	1,721
22	2,074	1,717
23	2,069	1,714
24	2,064	1,711
25	2,060	1,708
26	2,056	1,706
27	2,052	1,703
28	2,048	1,701
29	2,045	1,699
30	2,042	1,697
40	2,021	1,684
60	2,000	1,671
100	1,984	1,660
120	1,980	1,658
∞	1,960	1,645

Valorile critice ale statisticii U pentru testul Mann-Whitney pentru $\alpha = 0,05$ (bilateral), $\alpha = 0,025$ (unilateral)

n_1	n_2																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	0	1	2	3	4	4	5	6	7	8	9	11	11	12	12	13	14
5	-	-	-	1	2	3	5	6	7	8	9	10	11	12	13	14	15	16	17	18
6	-	-	-	2	3	5	6	8	9	11	12	13	14	15	16	17	18	19	20	21
7	-	-	-	3	5	6	8	10	11	12	14	15	16	17	18	19	20	21	22	23
8	-	-	-	4	6	8	10	11	13	14	16	17	18	19	20	21	22	23	24	25
9	-	-	-	4	7	9	11	13	15	16	18	19	20	21	22	23	24	25	26	27
10	-	-	-	5	8	11	13	15	17	18	20	21	22	23	24	25	26	27	28	29
11	-	-	-	6	9	12	14	16	18	20	22	23	24	25	26	27	28	29	30	31
12	-	-	-	7	11	13	16	18	20	22	24	25	26	27	28	29	30	31	32	33
13	-	-	-	8	12	15	17	20	22	24	26	27	28	29	30	31	32	33	34	35
14	-	-	-	9	13	16	19	22	24	26	28	29	30	31	32	33	34	35	36	37
15	-	-	-	10	14	17	21	24	26	28	30	31	33	34	35	36	37	38	39	40
16	-	-	6	11	15	18	22	25	28	30	32	33	35	36	37	38	39	40	41	42
17	-	-	6	11	16	20	23	27	29	32	34	35	37	38	39	40	41	42	43	44
18	-	-	7	12	17	21	24	28	31	33	35	37	38	40	41	42	43	44	45	46
19	-	-	7	13	18	22	26	29	33	35	37	39	40	42	43	44	45	46	47	48
20	-	-	8	13	19	23	27	31	34	37	39	41	42	44	45	46	47	48	49	50

Valorile critice ale statisticii *T* a testului Wilcoxon

n	Bilateral α=0,05	Unilateral α=0,05
5	-	0
6	-	2
7	2	3
8	3	5
9	5	8
10	8	10
11	10	13
12	13	17
13	17	21
14	21	25
15	25	30
16	29	35
17	34	41
18	40	47
19	46	53
20	52	60
21	58	67
22	65	75
23	73	83
24	81	91
25	89	100
26	98	110
27	107	119
28	116	130
29	126	140
30	137	151
35	195	213
40	264	286
45	343	371
50	434	466
60	648	690
70	907	960
80	1211	1276
90	1560	1638
100	1955	2045

Valorile critice pentru testul F_{max} (Hatley) pentru $\alpha = 0,05$

$n-1$	k										
	2	3	4	5	6	7	8	9	10	11	12
2	39,0	87,5	142	202	266	333	403	475	550	626	704
3	15,4	27,8	39,2	50,7	62,0	72,9	83,5	93,9	104	114	124
4	9,6	15,5	20,6	25,2	29,5	33,6	37,5	41,1	44,6	48,0	51,4
5	7,15	10,8	13,7	16,3	18,7	20,8	22,9	24,7	26,5	28,2	29,9
6	5,82	8,38	10,4	12,1	13,7	15,0	16,3	17,5	18,6	19,7	20,7
7	4,99	6,94	8,44	9,70	10,8	11,8	12,7	13,5	14,3	15,1	15,8
8	4,43	6,00	7,18	8,12	9,03	9,78	10,5	11,1	11,7	12,2	12,7
9	4,03	5,34	6,31	7,11	7,80	8,41	8,95	9,45	9,91	10,3	10,7
10	3,72	4,85	5,67	6,34	6,92	7,42	7,87	8,28	8,66	9,01	9,34
12	3,28	4,16	4,79	5,30	5,72	6,09	6,42	6,72	7,00	7,25	7,48
15	2,86	3,54	4,01	4,37	4,68	4,95	5,19	5,40	5,59	5,77	5,93
20	2,46	2,95	3,29	3,54	3,76	3,94	4,10	4,24	4,37	4,49	4,59
30	2,07	2,40	2,61	2,78	2,91	3,02	3,12	3,21	3,29	3,36	3,39
60	1,67	1,85	1,96	2,04	2,11	2,17	2,22	2,26	2,30	2,33	2,36
∞	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Valorile critice q pentru testul Tukey (α=0,05).

$n-1$	k								
	2	3	4	5	6	7	8	9	10
1	17,97	26,98	32,82	37,08	40,41	43,12	45,40	47,36	49,07
2	6,08	8,33	9,80	10,88	11,74	12,44	13,03	13,54	13,99
3	4,50	5,91	6,82	7,50	8,04	8,48	8,85	9,18	9,46
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99
6	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92
9	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15
17	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92
30	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65
120	2,80	3,36	3,68	3,92	4,10	4,24	4,36	4,47	4,56

Valorile critice ale distribuției *F* pentru α=0,05.

gl_{int}	gl_{ext}									
	1	2	3	4	5	6	7	8	9	10
2	18,5	19,0	19,2	19,3	19,4	19,4	19,4	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,77	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91

Valorile critice ale distribuției χ^2

Grade de libertate	α						
	0,99	0,95	0,90	0,10	0,05	0,01	0,001
1	0,000157	0,00393	0,0158	2,706	3,841	6,635	10,828
2	0,0201	0,103	0,211	4,605	5,991	9,210	13,816
3	0,115	0,352	0,584	6,251	7,815	11,345	16,266
4	0,297	0,711	1,064	7,779	9,488	13,277	18,467
5	0,554	1,145	1,610	9,236	11,070	15,086	20,515
6	0,872	1,635	2,204	10,645	12,592	16,812	22,458
7	1,239	2,167	2,833	12,017	14,067	18,475	24,322
8	1,646	2,733	3,490	13,362	15,507	20,090	26,124
9	2,088	3,325	4,168	14,684	16,919	21,666	27,877
10	2,558	3,940	4,865	15,987	18,307	23,209	29,588
11	3,053	4,575	5,578	17,275	19,675	24,725	31,264
12	3,571	5,226	6,304	18,549	21,026	26,217	32,909
13	4,107	5,892	7,042	19,812	22,362	27,688	34,528
14	4,660	6,571	7,790	21,064	23,685	29,141	36,123
15	5,229	7,261	8,547	22,307	24,996	30,578	37,697
16	5,812	7,962	9,312	23,542	26,296	32,000	39,252
17	6,408	8,672	10,085	24,769	27,587	33,409	40,790
18	7,015	9,390	10,865	25,989	28,869	34,805	42,312
19	7,633	10,117	11,651	27,204	30,144	36,191	43,820
20	8,260	10,851	12,443	28,412	31,410	37,566	45,315
21	8,897	11,591	13,240	29,615	32,671	38,932	46,797
22	9,542	12,338	14,041	30,813	33,924	40,289	48,268
23	10,196	13,091	14,848	32,007	35,172	41,638	49,728
24	10,856	13,848	15,659	33,196	36,415	42,980	51,179
25	11,524	14,611	16,473	34,382	37,652	44,314	52,620
26	12,198	15,379	17,292	35,563	38,885	45,642	54,052
27	12,879	16,151	18,114	36,741	40,113	46,963	55,476
28	13,565	16,928	18,939	37,916	41,337	48,278	56,892
29	14,256	17,708	19,768	39,087	42,557	49,588	58,301
30	14,953	18,493	20,599	40,256	43,773	50,892	59,703

ANEXA 3: FUNCȚII MICROSOFT OFFICE EXCEL

În general, în acest program se poate realiza orice calcul prin tastarea „=", urmat de numerele sau referințele căsuțelor (literă coloană_număr rând) ce conțin valorile de interes, între care se introduc semnele de operație ale formulei (+, -, *, /, sum, sqrt, power, abs, round, log – pentru detalii consultați meniu-ul Help al programului).

Modulul Data Analysis există în meniu „Tools" în versiunea 97-2003 sau „Data" în versiunea 2007. Dacă nu apare în meniu, modulul se poate activa astfel: pentru versiunea 97-2003 se alege „Tools" > „Add-Ins"; în versiunea 2007 se alege „Office Button" > „Excel Options" > „Add-Ins".

FUNCȚII SPECIALE ÎN ORDINEA SECȚIUNILOR DIN CARTE:

2.1. SCALE DE MĂSURARE ȘI TIPURI DE VARIABILE

=RANK(nr, ref, 1)+(COUNT(ref) + 1 – RANK(nr, ref, 0) – RANK(nr, ref, 1)/2
Returnează rangul unei valori (nr) dintr-un set de date (ref).

2.2. REPREZENTAREA DATELOR

=LOG10(nr)
Returnează valoarea logaritmului zecimal al unui număr (nr)
=MAX(ref)
Returnează valoarea maximă a unui set de date (ref).
=MIN(ref)
Returnează valoarea minimă a unui set de date (ref).
=FREQUENCY(ref 1, ref 2)
Returnează frecvența cumulată a valorilor unui set de date (ref 1) cuprinse între valoarea minimă și limita superioară a unei clase de frecvență (ref 2).

3.1. TENDINȚA CENTRALĂ

=MODE(ref)
Returnează valoarea modului unui set de date (ref).
=MEDIAN(ref)
Returnează valoarea medianei unui set de date (ref).
=AVERAGE(ref)

3.2. VARIABILITATEA

Returnează valoarea mediei aritmetice a unui set de date (ref).
=MAX(ref)
Returnează valoarea maximă a unui set de date (ref).
=MIN(ref)
Returnează valoarea minimă a unui set de date (ref).
=SUMSQ(ref)
Returnează suma pătratelor valorilor dintr-un set de date (ref).
=STDEV(ref)
Returnează valoarea deviației standard calculată ca radical din suma pătratelor împărțită la numărul gradelor de libertate $(n-1)$, a unui set de date (ref).
=STDEVP(ref)
Returnează valoarea deviației standard a populației calculată ca radical din suma pătratelor împărțită la numărul valorilor (n), a unui set de date (ref).
=VAR(ref)
Returnează valoarea variației calculată ca suma pătratelor împărțită la numărul gradelor de libertate $(n-1)$, a unui set de date (ref).
=VARA(ref)
Returnează valoarea variației populației calculată ca suma pătratelor împărțită la numărul valorilor (n), a unui set de date (ref).

Descrierea statistică a unei probe mai poate fi obținută prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**DESCRIPTIVE STATISTICS**" și completarea câmpurilor din fereastra de dialog urmată de bifarea „**Summary statistics**".

4. DISTRIBUŢII PROBABILISTICE

=BINOMDIST(x, k, p, cumulativ)

Returnează probabilitatea binomială asociată numărului rezultatelor de interes (x), în funcţie de numărul total de încercări (k), şi probabilitatea obţinerii unui anumit rezultat (p), exprimată necumulat (cumulative = false).

=POISSONDIST(x, $\bar{x}$, cumulativ)

Returnează probabilitatea Poisson asociată numărului rezultatelor de interes (x), în funcţie de valoarea medie ($\bar{x}$), exprimată necumulat (cumulative = false).

=NORMALIZE(x, $\bar{x}$, s)

Returnează valoarea z a unei anumite valori a unei variabile (x) în funcţie de media valorilor ($\bar{x}$) şi de deviaţia standard a acesteia (s).

=NORMSINV(prob)

Returnează scorul z corespunzător unei anumite probabilităţi (prob) exprimată ca o proporţie din distribuţia normală standard.

=NORMSDIST(z)

Returnează probabilitatea ca proporţie din distribuţia normală standard, din coada stângă a distribuţiei şi până la un anumit scor z.

=STANDARDIZE(nr, $\bar{x}$, s)

Returnează scorul z al unei valori (nr) în funcţie de medie ($\bar{x}$) şi deviaţie standard (s).

Transformarea datelor

=LOG10(x)

Returnează logaritmul zecimal (x') al unei valori ce trebuie transformate (x).

=POWER(10,x')

Returnează antilogaritmul zecimal al unei valori transformate x' prin logaritmare în baza10.

=LN(x)

Returnează logaritmul natural (x') al unei valori ce trebuie transformate (x).

=EXP(x')

Returnează antilogaritmul natural al unei valori transformate x' prin logaritmare în baza e.

=LOG(x, bază)
Returnează logaritmul într-o bază specificată (bază) al unei valori ce trebuie transformate (x).
=POWER(bază,x')
Returnează antilogaritmul unei valori transformate x' prin logaritmare într-o bază specificată (bază).
=ASINH(x)
Returnează valoarea transformată prin funcția arcsinh a unei valori x ce trebuie transformată.
=SINH(x')
Returnează valoarea transformării inverse a unei valori x' transformată cu funcția arcsinh.
=SQRT(x)
Returnează valoarea transformată prin extragerea radicalului dintr-o valoare x ce trebuie transformată.
=POWER(x',2)
Returnează valoarea transformării inverse a unei valori x' transformată prin extragerea radicalului.
=DEGREES(ASIN(SQRT(x)))
Returnează valoarea transformării unei proporții ($0 \leq x \leq 1$) cu ajutorul funcției arcsin.
=POWER(SIN(RADIANS(x')), 2)
Returnează o proporție (x) prin transformarea inversă a unei valori obținută cu ajutorul funcției arcsin (x').

5.1. ESTIMAREA MEDIEI POPULAȚIONALE

Estimarea mediei populaționale poate fi obținută prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**DESCRIPTIVE STATISTICS**" și completarea câmpurilor din fereastra de dialog urmată de bifarea „**Summary statistics**" și „**Confidence Level for Mean**". Ultimul rând din tabelul de rezultate conține valoarea produsului dintre eroarea standard a mediei și valoarea critică a distribuției Student pentru α (valoarea implicită este de 0,05) și $n-1$ grade de libertate.

=TINV(prob, gl)

Returnează valoarea critică t (Student) pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $n-1$).

6. TESTAREA UNEI IPOTEZE PRIVIND MEDIA UNEI SINGURE POPULAȚII

=TINV(prob, gl)

Returnează valoarea critică t (Student) pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $n-1$).

=TDIST(*t*, gl, cozi)

Returnează probabilitatea asociată unei valori t (Student) pentru un anumit număr de grade de libertate (gl = $n-1$) și în funcție de varianta testului (cozi: 1 – unilateral, 2 – bilateral).

7.1.1. Testul t (Student) pentru probe independente

=TINV(prob, gl)

Returnează valoarea critică t (Student) pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $n-1$).

=TDIST(*t*, gl, cozi)

Returnează probabilitatea asociată unei valori t (Student) pentru un anumit număr de grade de libertate (gl = $n-1$) și în funcție de varianta testului (cozi: 1 – unilateral, 2 – bilateral).

Testul Student pentru probe independente se poate realiza prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**t-Test: Two-Sample Assuming Unequal Variances**" și completarea câmpurilor din fereastra de dialog. În câmpul „**Hypothesized Mean Diference**" se completează valoarea 0.

7.2.1. Testul t (Student) pentru perechi de observații

=TINV(prob, gl)

Returnează valoarea critică t (Student) pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $n-1$).

=TDIST(t, gl, cozi)

Returnează probabilitatea asociată unei valori t (Student) pentru un anumit număr de grade de libertate (gl = $n-1$) și în funcție de varianta testului (cozi: 1 – unilateral, 2 – bilateral).

Testul Student pentru perechi de observații se poate realiza prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**t-Test: Paired Two Sample for Means**" și completarea câmpurilor din fereastra de dialog. În câmpul „**Hypothesized Mean Diference**" se completează valoarea 0.

8.1.1. Testarea omogenității variației interne

=VAR(ref)

Returnează valoarea variației calculată ca suma pătratelor împărțită la gradele de libertate ($n-1$), a unui set de date (ref).

=LN(nr)

Returnează valoarea logaritmului natural a unui număr (nr).

=CHIINV(prob, gl)

Returnează valoarea critică χ^2 pentru un anumit nivel de semnificație (prob = α), un anumit număr de grade de libertate (gl = $k-1$).

=CHIDIST(χ^2, gl)

Returnează probabilitatea asociată unei valori χ^2 pentru un anumit număr de grade de libertate (gl = $k-1$).

8.2.1. ANOVA model unifactorial

=FINV(prob, glext, glint)

Returnează valoarea critică F pentru un anumit nivel de semnificație (prob = α), un anumit număr de grade de libertate externe (glext = $k-1$) și un anumit număr de grade de libertate interne (glint = n_t-k).

=FDIST(*F*, glext, glint)

Returnează probabilitatea asociată unei valori F pentru un anumit număr de grade de libertate externe (glext = $k-1$) și un anumit număr de grade de libertate interne (glint = n_t-k).

Testul ANOVA model unifactorial se poate realiza prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**Anova: Single Factor**" și completarea câmpurilor din fereastra de dialog.

8.2.2. ANOVA Kruskal-Wallis

=CHIINV(prob, gl)

Returnează valoarea critică χ^2 pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $k-1$).

=CHIDIST(χ^2, gl)

Returnează probabilitatea asociată unei valori χ^2 pentru un anumit număr de grade de libertate (gl = $k-1$).

8.2.3. ANOVA model bifactorial fără replicare

=FINV(prob, glext, glint)

Returnează valoarea critică F pentru un anumit nivel de semnificație (prob = α), un anumit număr de grade de libertate externe (glext = $c-1$ sau glext = $r-1$) și un anumit număr de grade de libertate interne (glint = $(c-1)(r-1) = n_t - c - r + 1$).

=FDIST(*F*, glext, glint)

Returnează probabilitatea asociată unei valori F pentru un anumit număr de grade de libertate externe (glext = $c-1$ sau glext = $r-1$) și un anumit număr de grade de libertate interne (gl = $(c-1)(r-1) = n_t - c - r + 1$).

Testul ANOVA model bifactorial cu o singură observație în celulă se poate realiza prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**Anova: Two Factor Without Replication**" și completarea câmpurilor din fereastra de dialog.

8.2.4. ANOVA Friedman

=CHIINV(prob, gl)

Returnează valoarea critică χ^2 pentru un anumit nivel de semnificație (prob = α), un anumit număr de grade de libertate (gl = $c - 1$).

=CHIDIST(χ^2, gl)

Returnează probabilitatea asociată unei valori χ^2 pentru un anumit număr de grade de libertate (gl = $c - 1$).

8.2.5. ANOVA model bifactorial cu replicare

=FINV(prob, glext, glint)

Returnează valoarea critică *F* pentru un anumit nivel de semnificație (prob = α), un anumit număr de grade de libertate externe (glext = $c - 1$ sau glext = $r - 1$ sau glext = (c–1)(r–1)) și un anumit număr de grade de libertate interne (glint = $n_t - cr$).

=FDIST(*F*, glext, glint)

Returnează probabilitatea asociată unei valori *F* pentru un anumit număr de grade de libertate externe (glext = $c - 1$ sau glext = $r - 1$ sau glext = (c–1)(r–1)) și un anumit număr de grade de libertate interne (glint = $n_t - cr$).

Testul ANOVA model bifactorial cu o singură observație în celulă se poate realiza prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**Anova: Two Factor With Replication**" și completarea câmpurilor din fereastra de dialog.

9.1. Analiza Corelației

=CORREL(ref *x*, ref *y*)

Returnează valoarea coeficientului de corelație parametrică Pearson dintre valorile variabilei x (ref *x*) și valorile variabilei y (ref *y*).

=TINV(prob, gl)

Returnează valoarea critică t (Student) pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $n-2$).

=TDIST(t, gl, cozi)

Returnează probabilitatea asociată unei valori t (Student) pentru un anumit număr de grade de libertate (gl = $n-2$) și în funcție de varianta testului (cozi: 1 – unilateral, 2 – bilateral).

Coeficientul de corelație parametrică Pearson se poate realiza prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**Correlation**" și completarea câmpurilor din fereastra de dialog.

9.2. Analiza Regresiei

Analiza regresiei se poate realiza prin apelarea meniului „**DATA**" > „**DATA ANALYSIS**" > „**Regression**" și completarea câmpurilor din fereastra de dialog.

10.1. Testul χ^2 de concordanță

=CHIINV(prob, gl)

Returnează valoarea critică χ^2 pentru un anumit nivel de semnificație (prob = α), un anumit număr de grade de libertate (gl = $k-m-1$).

=CHIDIST(χ^2, gl)

Returnează probabilitatea asociată unei valori χ^2 pentru un anumit număr de grade de libertate (gl = $k-m-1$).

=BINOMDIST(x, k, p, cumulativ)

Returnează probabilitatea binomială asociată numărului rezultatelor de interes (x), în funcție de a numărului total de încercări (k), și a probabilității obținerii unui anumit rezultat (p), exprimată necumulat (cumulative = false).

=POISSONDIST(x, $\bar{x}$, cumulativ)

Returnează probabilitatea Poisson asociată numărului rezultatelor de interes (x), în funcție de valoarea medie ($\bar{x}$), exprimată necumulat (cumulative = false).

10.2. Testul χ^2 de asociere

=CHIINV(prob, gl)

Returnează valoarea critică χ^2 pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $(c-1)(r-1)$).

=CHIDIST(χ^2, gl)

Returnează probabilitatea asociată unei valori χ^2 pentru un anumit număr de grade de libertate (gl = $(c-1)(r-1)$).

10.3. Testul exact al lui Fisher

=FACT(x)

Returnează valoarea $x!$.

10.4. Testul McNemar

=CHIINV(prob, gl)

Returnează valoarea critică χ^2 pentru un anumit nivel de semnificație (prob = α) și un anumit număr de grade de libertate (gl = $(c-1)(r-1)$).

=CHIDIST(χ^2, gl)

Returnează probabilitatea asociată unei valori χ^2 pentru un anumit număr de grade de libertate (gl = $(c-1)(r-1)$).

=BINOMDIST(x, k, p, cumulativ)

Returnează probabilitatea binomială asociată numărului rezultatelor de interes (x), în funcție de a numărului total de încercări (k), și a probabilității obținerii unui anumit rezultat (p), exprimată necumulat (cumulative = false).

INDEX ALFABETIC

U

V

W

Z

TIPARUL EXECUTAT LA
IMPRIMERIA EDITURII UNIVERSITĂȚII
„ALEXANDRU IOAN CUZA" DIN IAȘI
700511 Iaşi, Păcurari 9, tel./fax 0232 314947
Format: 70×100/16
Apărut: 2009
Comanda: 5

Informaţii şi comenzi:
www.editura.uaic.ro
editura@uaic.ro

Ştefan R. Zamfirescu
Oana Zamfirescu

ELEMENTE DE STATISTICĂ APLICATE ÎN ECOLOGIE

În prezent, valorificarea investigaţiilor ecologice nu poate fi concepută fără o analiză statistică a datelor, fără aşa-numita asigurare statistică ce redă măsura în care concluziile investigaţiilor ar putea fi reale. Analiza statistică a datelor, fără a fi un scop în sine al demersului ştiinţific în ecologie, reprezintă o unealtă ce permite o mai bună comprehensiune şi prezentare a informaţiei conţinute de rezultatele cercetărilor. În prezent, prelucrarea statistică a datelor este facilitată de utilizarea computerului şi a programelor dedicate.

Prezenta lucrare încearcă să ofere o bază conceptuală pentru cei care se dedică cercetărilor cu caracter ecologic. Din acest motiv, pe parcursul capitolelor apar numeroase exemple inspirate din cercetări ecologice. Lucrarea poate fi însă de ajutor şi pentru analiza datelor rezultate în urma investigaţiilor din diverse ramuri ale biologiei sau ştiinţe conexe.

www.editura.uaic.ro

9789737 033895

Editura Universităţii „Alexandru Ioan Cuza" Iaşi